MULTIPLE IMPUTATION TO CORRECT FOR MEASUREMENT ERROR:

APPLICATION TO CHRONIC DISEASE CASE ASCERTAINMENT IN ADMINISTRATIVE HEALTH DATABASES

A Thesis Submitted to the College of

Graduate Studies and Research

In Partial Fulfillment of the Requirements

For the Degree of Master of Science

In the Collaborative Graduate Program in Biostatistics

University of Saskatchewan

Saskatoon

By

XUE YAO

PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

Requests for permission to copy or to make other uses of materials in this dissertation in whole or part should be addressed to:

Graduate Chair, Collaborative Biostatistics Program

School of Public Health

University of Saskatchewan

107 Wiggins Road

Saskatoon, Saskatchewan

S7N 5E5 Canada

ABSTRACT

Diagnosis codes in administrative health databases (AHDs) are commonly used to ascertain chronic disease cases for research and surveillance. Low sensitivity of diagnosis codes has been demonstrated in many studies that validate AHDs against a gold standard data source in which the true disease status is known. This will result in misclassification of disease status, which can lead to biased prevalence estimates and loss of power to detect associations between diseases status and health outcomes. Model-based case detection algorithms in combination with multiple imputation (MI) methods in validation dataset/main dataset designs could be used to correct for misclassification of chronic disease status in AHDs. Under this approach, a predictive model of disease status (e.g., logistic model) is constructed in the validation dataset, the model parameters are estimated and MI methods are used to impute true disease status in the main dataset. This research considered scenarios that the misclassification of the observed disease status is independent of disease predictors and dependent on disease predictors. When the misclassification of the observed disease status is independent of disease predictors, the MI methods based on Frequentist logistic model (with and without bias correction) and Bayesian logistic model were compared. And when the misclassification of the observed disease status is dependent on disease predictors, the MI based on Frequentist logistic model with different variables as covariates were compared. Monte Carlo techniques were used to investigate the effects of the following data and model characteristics on bias and error in chronic disease prevalence estimates from AHDs: sensitivity of observed disease status based on diagnosis codes, size of the validation dataset, number of imputations, and the magnitude of measurement error in covariates of the predictive model. Relative bias, root mean squared error and coverage of 95% confidence interval were used to measure the performance. Without bias correction, the Bayesian MI model has lower RMSE than the Frequentist MI model. And the Frequentist MI model with bias correction is demonstrated via a simulation study to have superior performance to Bayesian MI model and the Frequentist MI model without bias correction. The results indicate that MI works well for measurement error correction if the missing true values are not missing not at random no matter whether the observed disease diagnosis is dependent on other disease predictors or not. Increasing the size of the validation dataset can improve the performance of MI better than increasing the number of imputations.

ACKNOWLEDGMENTS

I would like to first gratefully and sincerely thank my supervisor Dr. Lisa Lix for her insightful supervision and constant support and encouragement during the program. Her wisdom, knowledge and commitment to the highest standards inspired and motivated me as well as influenced me profoundly in my life. I am so proud to be one of her students.

I truly and deeply appreciate the comments provided by my thesis committee member Dr. Juxin Liu. Her criticisms and suggestions in the theory of statistics have improved the statistical methodology in this thesis. My gratitude also goes to the current Chair of the Collaborative Biostatistics Program, Dr. Mik Bickis and my external examiner Dr. Nazmi Sari for offering constructive advice.

I also acknowledge the Canadian Institutes of Health Research, Saskatchewan Health Research Foundation, and University of Saskatchewan College of Graduate Study and Research for the funding provided for this project and my graduate study.

Thanks to all past and current members in the Population Health Data Laboratory at the University of Saskatchewan for sharing their knowledge and enthusiasm.

I cherish all the friends and their support through my graduate study. Not enough can be said of the patience, understanding and support of my dearest friend Chaojie Lin. I am grateful to her for standing by me since I applied for this program.

Last but not least, I would like to express my deepest gratitude to my beloved parents and my brother for their endless love and support in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AHD	Administrative Health Database
CART	Classification and Regression Trees
GFR	Glomerular Filtration Rate
ICD	International Classification of Diseases
MAR	Missing at Random
MCAR	Missing Completely at Random
MCMC	Markov Chain Monte Carlo
MI	Multiple Imputation
ML	Maximum Likelihood
MNAR	Missing not at Random
MSE	Mean Squared Error
RC	Regression Calibration
RMSE	Root Mean Squared Error
SE	Standard Error

CHAPTER 1. INTRODUCTION

1.1 Background

Population-based administrative health databases (AHDs), including hospital records and physician claims, are widely used for chronic disease research and surveillance[1-4]. In AHDs, diagnostic information is usually based on the International Classification of Diseases (ICD) codes, which was developed by the World Health Organization. The presence or absence of a diagnosis code in AHD is used to ascertain disease status (i.e., disease presence/absence). Ascertained disease status is used to estimate prevalence and incidence in surveillance studies. As well, the association between disease presence and health outcomes, such as hospitalization or death, might also be investigated using AHDs. Accurate disease case ascertainment is important not only for obtaining accurate prevalence estimates but also for producing unbiased epidemiologic and clinical studies about disease outcomes.

AHD was originally developed for health system management and physician remuneration. The accuracy of diagnoses in AHDs for research has been questioned. Validation studies, in which AHDs are linked, via a unique personal identifier, to a 'gold standard' data source in which the true disease status is known, have been used to assess the accuracy of diagnoses recorded in AHDs. In general, these studies have demonstrated high specificity but low sensitivity of diagnoses[5-7], although sensitivity and specificity will vary across chronic diseases[8]. Sensitivity is defined as the probability of correctly identifying disease individuals from all the individuals who truly have the disease, while specificity is defined as the probability of correctly identifying non-disease individuals from amongst those who truly do not have the disease[9].

Inaccuracies in diagnosis codes in AHDs lead to misclassification of observed disease status. In particular, low sensitivity and perfect specificity will result in under-reporting of disease prevalence[10,11]. Misclassification of observed disease status can also attenuate the association between disease presence as a response variable and risk factors, or the association between disease presence as an explanatory variable and health outcomes[12]. Model-based case-detection algorithms, which can improve sensitivity of the observed disease status without loss of specificity, have been proposed for AHDs as well as for other related problems[11,13]. Model-based

algorithms develop a predictive model for chronic disease status in a validation dataset, that is a dataset in which the true disease status is known and the observed disease status (i.e., ascertained based on diagnosis codes) is also captured. The model is then applied to another dataset, known as the main dataset, which only contains observed disease status, to predict true disease status. The model that is fitted to the validation dataset predicts true disease status from disease predictor variables; these predictor variables are also found in the main dataset. Examples of disease predictor variables could include the presence of prescription drug treatments, comorbid conditions, severity of illness, and demographic variables. True disease status can be predicted using a parametric model, such as logistic regression, or a non-parametric model, such as classification and regression tree (CART) analyses. Model-based case-detection algorithms have been shown to have better discriminative performance and lower prediction error than case-detection algorithms based on diagnosis codes alone. However, there are several issues that arise in the development of the predictive model. First, some disease predictors, particularly those based on ICD codes, may also contain measurement/misclassification error[14]. Second, there are several approaches that can be used to construct model-based case-detection algorithms and it is not clear which approach should be preferred.

A number of studies have proposed treating measurement error as a missing-data problem[15-17]. In particular, multiple imputation (MI) methods have been proposed to accurately estimate true disease status and its variation in main/validation dataset designs. MI is a flexible technique that has been applied to problems of missing data[18,19] and confidential data[20], as well as measurement error[15,21,22]. A primary advantage of using a MI method is that it is relatively straightforward to implement, which is beneficial for applied researchers. MI methods based on both Frequentist and Bayesian paradigms have been proposed and applied[17,23,24], but few, if any studies have compared these methods. As well, there has been limited investigation about the characteristics of the study design and MI methodology that may influence the performance of model-based case-detection algorithms. Such studies could help researchers take maximum advantage of AHDs for chronic disease research and surveillance.

1.2 Purpose and Objectives

The purpose of this research is to investigate the performance of MI model-based case detection methods for estimating chronic disease prevalence in validation/main datasets designs. The objectives are:

1. To investigate MI methods using Frequentist and Bayesian logistic regression model as predictive model;
2. To examine characteristics of model-based case-detection algorithms that may influence the performance of MI methods, including sensitivity of AHD diagnosis codes, and type and magnitude of measurement error in the predictors of disease status; and
3. To investigate the effects of size of the validation dataset and number of imputations on the performance of MI methods.

CHAPTER 2. LITERATURE REVIEW

The literature review covers the following topics: measurement error, validity of diagnosis codes in AHD, measurement error correction methods, MI methods, and software to implement MI methods.

2.1 Types of Measurement Error

Measurement error is defined as the difference between the observed value of a variable and the true value of a variable. Measurement error in categorical variables is usually referred to as misclassification. When a categorical variable is subject to misclassification, sensitivity and specificity are used to quantify the accuracy of the measurement.

Measurement error can be defined using both Classical and Berkson models[16]. For the Classical measurement error model, the measurement measures the truth with additive error, usually with homoscedastic variance. For example, in feeding studies, researchers have posited that protein biomarkers will capture true protein intake with added constant variability[25,26]. The Classical model is defined as $W = X + \varepsilon$, where W denotes the observed variable, X denotes the true variable and ε is the error term. In the Berkson measurement error model, the true value contains more variability than the measured value. For example, in the Hanford Thyroid Disease Study, it is thought that the true thyroid dose equals to the measured dose plus error[27]. The Berkson model is $X = W + \varepsilon$, indicating that the variability of the true variable (X) equals the sum of the variability of the observed variable (W) and error (ε). If a variable is measured uniquely for all individuals, measurement error is likely to follow the Classical model. For example, if people fill out a self-report questionnaire about their health conditions or if they get a blood pressure measurement, then the errors will likely follow the Classical model. However, if all individuals in a stratum are given the same value of the measure but the true value is specific to an individual, then the Berkson model is a reasonable choice. For a given measurement error variance, measurement error in the Classical model results in greater loss of power than measurement error in the Berkson model[16].

Measurement error can also be characterized as differential or non-differential. Non-differential measurement error means that the observed covariate contains no information about the response given the true covariate. Measurement error is non-differential if the observed

covariate is independent of the response, conditional on the values of the true covariate. Otherwise, the measurement error is differential. For example, in the Framingham study, which is a large cohort study about the cardiovascular disease predictors, one response of interest is the presence of coronary heart disease; the main predictor of interest is systolic blood pressure which is not possible to measure directly[28,29]. Only blood pressure measurements observed during a clinic visit are available. Since given the long term systolic blood pressure one visit's measured blood pressure provides no information to the coronary heart disease, the measurement error is non-differential[30]. In AHDs when the interested response is the disease status, if the probabilities of mismeasurement of disease predictors are assumed to be the same for persons with the disease as for persons without the disease, the measurement error is non-differential. However, measurement error is differential if the probabilities differ for persons with and without disease. In case-control studies, the response is observed first and then subsequent follow-up ascertains the predictors of the response, which is known as recall bias, so the measurement of the predictors usually depend on the response and are known as differential measurement. And differential measurement error typically occurs in retrospective studies. If the response is measured after the covariates are measured, which is typical of cohort studies, measurement error in the covariates tends to be non-differential. Carroll et al.[16] concluded that additive, unbiased, homoscedastic response measurement error in linear or nonlinear regression will result in increased variability of the fitted models and decreased power to detect effects of covariates. Misclassification of the response not only masks the features of the data but also results in bias in the parameter estimates.

2.2 Measurement Error in AHD

A number of studies have investigated the validity of diagnosis codes in AHD for ascertaining cases of chronic disease.[31-35] Under reporting of disease cases, in which the observed value has high specificity (i.e., close to 1.00) but low sensitivity (i.e., less than 0.95), appears to be a common problem in AHDs. Data sources that have been used to validate administrative databases include medical records, patient or physician surveys, and clinical laboratory test results[6,36-38]. For example, Rector et al.[3] investigated the validity of diagnosis codes for identifying cases of hypertension, heart failure, chronic lung disease, arthritis, glaucoma, and diabetes using survey data as the gold standard. For all six conditions the

specificities of diagnosis-based case-detection algorithms were greater than 0.95 but sensitivities were rarely higher than 0.90. A study about hypertension found that a diagnosis-based case-detection algorithm had a sensitivity of 0.73 and a specificity of 0.95 when medical records were used as the gold standard and a sensitivity of 0.64 and a specificity of 0.94 when self-reported survey data were used as the gold standard[6]. Osteoarthritis is another disease that is prone to misclassification in AHDs; it has been estimated that the sensitivity of diagnosis-based case-detection algorithms is between 0.70 and 0.80, while specificity is between 0.90 and 1.00[12].

Diagnoses for comorbid conditions are also prone to misclassification in AHDs. Comorbid conditions are co-occurring diseases that are related to the primary disease. Validation studies based on chart data have shown that comorbidities are likely to be underestimated in AHDs[37]. As well, the type of AHD that is used to ascertain comorbidities may also influence sensitivity. Klabunde et al. demonstrated that in cohorts of elderly prostate and breast cancer patients the proportion of patients identified with comorbid conditions increased to as much as 25 percent by using physician claims data to identify diagnoses instead of hospital records. The latter data source could detect less than 10 percent of the true comorbid conditions[39]. However, physician claims data were still prone to misclassification of comorbid conditions. A Canadian study that compared chart data and administrative data revealed that the latter generally underestimate individual comorbidities[37]. Another study that undertook a chart review in 817 hospitalized patients receiving percutaneous coronary interventions also found that administrative data tended to underestimate the prevalence of some comorbidity[40]. Therefore, if comorbid conditions are used as disease predictor variables in model-based case-detection algorithms, they may induce bias in the predictive model.

The effect of error in the response variable on bias in the covariate effect estimates depends on the sensitivity and specificity of the observed response; when both sensitivity and specificity are low, the estimate of the covariate effect will be biased downward[41]. For measurement error in a continuous covariate, the amount of error also impacts on the performance of measurement error correction methods[16]. The characteristics of the measurement error or misclassification deserve more investigation.

2.3 Measurement Error Correction Methods

Measurement error correction methods have primarily been developed for the case of measurement error or misclassification in model covariates[22,42-47]. For example, one approach for measurement error in covariates is the regression calibration (RC) model, which uses information from a validation dataset to analyze the association between the true and observed variables and then corrects the estimate of association in the main dataset[48,49]. The RC method is popular because it is a simple approach to use. However, it can be problematic in situations where the covariates have skewed distributions[50]. When the covariate is analyzed on a categorical scale, the RC method will result in similar estimates of the effects of covariates as the estimates without measurement error correction[51].

A second approach to address measurement error in covariates is the maximum-likelihood (ML) method, which maximizes the likelihood function as if the true values were observed based on the assumed measurement error model (i.e., Classical or Berkson model)[52-55]. Firstly, the likelihood function of response given the true covariates is constructed. And then depending on whether the error model is assumed to be Classical or Berkson, the likelihood analysis is defined based on the joint distribution of the response, true covariates and observed covariates. This method can derive reliable likelihood-based confidence intervals in nonlinear models but model misspecification is a serious limitation. Pepe demonstrated that the ML parameter estimates are not robust to misspecification of the error model[56]. As well, ML estimation of misspecified models have been shown to result in invalid inference[57].

MI methods has been investigated to correct for measurement error in covariates and compared with both RC and ML methods[58]. MI and RC methods were compared for estimating the hazard ratios in a simulation study based on a real study about end-stage renal disease[15]. The variable of interest was the glomerular filtration rate (GFR), which is considered to be an error prone measure of renal function. In the simulation study, sensitivities of 0.7 or 0.9 and specificities of 0.7 or 0.9 were investigated for the GFR. The MI method produced unbiased estimates of the hazard ratio and had approximately correct coverage of the 95% confidence interval. When the sensitivity and specificity of the observed GFR was low the MI method was more powerful than the RC method. Other studies have shown that for validation dataset/main

dataset designs, the MI method results in smaller bias and similar error rates when compared to the ML method[58].

Methods for inference and estimation in the presence of an error-prone binary response variable have also been proposed[56,59-61]. However, most of these methods assume the covariates are measured without error. One exception is the method by Mallick and Gelfand; they proposed a Bayesian approach based on a semiparametric generalized linear model, which can accommodate measurement error in both the response variable and the covariates[62]. Their Bayesian model involves fitting three separate regression models for measurement error in the covariates, the response variable, and the relationship between the response variable and covariates. For each regression model, a semiparametric generalized linear model was introduced utilizing an unknown monotonic function (nonparametric links) along with parametric regression coefficients. However, only a Poisson regression model for the link function of the response and covariates was illustrated by simulation. Different link functions were specified to investigate the ability to modify misspecified true functions of the semiparametric models. The results demonstrated the large sensitivity of the method to the link function specification.

Chakraborty and Banerjee[63] considered bivariate regression models containing one binary response and one continuous response when the covariate was error prone and the binary response was misclassified. The covariate was assumed to follow a normal distribution with a Berkson error model and non-differential measurement error. The responses were generated from bivariate normal distribution, and the binary response was obtained by dichotomizing a continuous variable. A simulation study was undertaken to investigate the effects of measurement error and/or misclassification on the model parameter estimates based on different choices of the additive variance of the observed covariate and probability of misclassification. Four models were compared, including a naïve model (i.e., without consideration of measurement error and misclassification), a model incorporating classification error, a model incorporating measurement error, and a model incorporating both classification error and measurement errors. Misclassification attenuated the estimated correlation between the response variables, while measurement error inflated the coefficient estimates. The attenuating effect of measurement error in the error prone covariate on the estimated covariate coefficients became larger as the additive variance of the measurement error increased, when the binary response was

not misclassified. When the additive variance was 0.5, the estimate of the correlation between the response variables whose true value was 0.60 was estimated to be 0.73, but when the additive variance increased to 1.0, the estimated correlation increased to 0.80. When measurement error was absent, misclassification of the binary response also attenuated the coefficient estimates of the covariate and the bias became larger as the misclassification probability increased. For instance, when the probability of misclassification was 0.01 (i.e., the sensitivity and specificity were 0.99), the estimate of the coefficient of the covariate equaling to 1.00 was 0.73; when the probability of misclassification was 0.05 (i.e., the sensitivity and specificity were 0.95), the estimate of the coefficient of the covariate decreased to 0.55. The study demonstrated that if both measurement error and misclassification exist in the data, the attenuation of the estimate of the covariate's coefficient became pronounced with increase in the value of the misclassification error rates as well as measurement error variance. The measurement errors result in attenuation of the estimate of the correlation between response variables, and the misclassification results in inflation of the estimate of the correlation between response variables.

2.4 MI Methods

MI methods have been used to address nonresponse in large survey datasets, such as the US National Health and Nutrition Examination Survey[64], the US Survey of Consumer Finance[65], the US National Health Interview Survey[66], and the Cancer Care Outcomes Research and Surveillance Consortium[67]. MI methods have also been used to handle missing data in non-survey contexts[24]. As noted in the previous section, MI methods were applied to observational health care outcomes data to address measurement error in binary treatment variables[60].

One consideration in adopting a MI method is the mechanism by which observations are missing. The data may be missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). If missingness of the observations does not depend on the actual values, regardless of whether they are missing or observed, the missing data are MCAR. On the other hand, if the missingness mechanism is MAR, this means that the missingness depends only on the observed values and not on the missing values. Under the MAR assumption, the missing data may depend on the data itself, but only indirectly through relationships with observed values. If the missingness mechanism is MNAR, missingness depends on the missing

values[68]. Research has demonstrated that even if the missingness mechanism is MNAR, the MI method will not always result in biased regression parameter estimates and standard errors[69]; however the size of the bias is dependent on the amount of missing data. Moreover, some studies have applied the MAR missingness mechanism as a plausible assumption [60,70]. For the validation/main datasets design, the sampling frame of the validation subjects determines the missingness mechanism. If the individuals in the validation dataset are a random sample from the population, the missing values in the main dataset are MCAR. If individuals' presence in the validation dataset is dependent on variables that are observed in both the validation and main datasets, the missing values in the main dataset are MAR. And if individuals' presence in the validation dataset is dependent on variables that are not observed (i.e., missing) in the main dataset, the missing values in the main dataset are MNAR.

Yucel and Zaslavsky[71] applied imputation methods for binary treatment variables with measurement error in administrative data. In their study, the administrative database was a cancer registry that collected data on treatment and survival for all incident colorectal cancer cases in California. The validation dataset was a dataset collected by surveying physicians or reviewing office records, which covered 74% patients. The MAR missingness mechanism was assumed, because treatment information was obtained less often from physicians' surveys or office records for patients 75 years of age and older, unmarried patients, nonrural patients, those in low-volume hospitals or hospitals with radiation facilities, and those living in San Francisco/Oakland than in San Jose/Monterey or Sacramento.

MI accounts for the uncertainty introduced by the imputation process and reduces potential bias due to systematic differences between the observed and missing data[68]. Compared with a single imputation approach, MI results in accurate estimates of the confidence intervals. Messer and Natarajan studied the ML, MI, and RC methods for adjustment in the covariates and suggested that MI methods can be an appropriate approach to impute unobserved values of both the response and covariates[58].

Rubin proposed MI based on Bayesian theory[24]. MI methods derived from a Frequentist approach have also been evaluated[72]. The Bayesian and Frequentist approaches handle uncertainty in model parameters differently. Under the Bayesian approach, the model parameters are treated as random variables whose prior distributions can be specified based on experience

and other sources. The likelihood and prior distribution determine the posterior distribution of the parameter[73]. MI methods under the Bayesian approach draw values of the unobserved variables by sampling from the posterior distribution, given the observed values and the other parameters[23]. Under the Frequentist approach, the values of the parameters are fixed and the MI method is used to obtain multiple draws of the plausible values from the asymptotic distribution of the parameters as estimated from a predictive model. Thus, when imputing the unobserved true covariate, the Bayesian approach conditions on the response as well as the observed measurement of the true covariate(s), while the Frequentist approach only depends on the observed measurement. In addition, the Frequentist inference is based on asymptotic approximations; the Bayesian approach converges stochastically to a posterior distribution that is exact, regardless of sample size[72]. To implement MI in the Bayesian approach when the likelihood function is complicated to derive, the sampling-based Markov Chain Monte Carlo (MCMC) method has been proposed[23]. A Markov chain is a sequence of random variables in which each element's value depends only on the value of the previous element and the chain converges to a stationary posterior distribution.

In addition, the size of the validation dataset as a proportion of the entire population will have an influence on inference using the MI method. The MI method proposed by Rubin works best if the missingness rate is less than forty percent. Consequently, the efficiency and consistency of MI methods when the missingness rate is higher than this requires further consideration. The relative efficiency of an estimate based on *K* imputations compared to an estimate based on an infinite number of imputations is $\left(1+\frac{\lambda}{K}\right)^{-1}$, where λ is the rate of missing data[68]. As few as three to 10 imputations will produce efficient results[74]. However, one simulation study about a simple regression model showed that when λ is held constant, as the number of imputations increased from three to 100 the values of mean squared error (MSE) and standard error (SE) decreased and power increased. For instance, when the proportion of missing information was 0.30 the MSE, SE and power for three imputations were 1.31, 0.04, and 0.69, respectively. In contrast, the MSE, SE and power based on 100 imputations were 1.20, 0.04, and 0.79, respectively. When the fraction of missing information was 0.90, the MSE, SE and power for $K = 3$ were 1.67, 0.04, and 0.39, while the corresponding values for 100 imputations were

1.21, 0.04, and 0.78, respectively. Consequently, the authors recommended using 40 imputations when λ = 0.70 and 100 imputations when λ = 0.90[75].

2.5 MI in Practice

Software packages that implement MI methods include SAS, S-PLUS, Stata, R and MICE (Multivariate Imputation by Chained Equations). In SAS, PROC MI and PROC MIANALYZE are available to implement MI methods. However, most MI methods in SAS are available for monotone missing pattern data, when the pattern of missingness is arbitrary, only continuous variables can be imputed using PROC MI[76]. There are also a few SAS macros for MI using sequences of regression models or distance-aided selection of donors[77]. The MICE method has also been implemented as a S-PLUS library and an R package as well as in Stata. Overall, these programs provide an easy to use environment for applying MI. However, the restrictions of assumptions still limit the application of MI in specific study[78].

2.6 Summary of Literature Review

Previous research has demonstrated that measurement error exists in chronic disease outcomes and comorbidities defined from diagnosis codes in AHDs. This measurement error is primarily due to low sensitivity of the diagnoses in AHDs. While a few procedures have been proposed to address measurement error in AHDs, developing a predictive model that uses MI methods to impute disease status is appealing because this approach should be relatively straightforward to implement and interpret. MI methods based on Frequentist or Bayesian approaches have been proposed and applied in real datasets, but the performance of both approaches has not been compared in previous research. Factors that have been shown to directly or indirectly influence the performance of the MI method for measurement error correction have been investigated, but few studies have simultaneously investigated these factors when both the response variable and the covariates are error prone. Factors that have been investigated in previous research that may influence the performance of MI measurement error correction methods included the sensitivity of the error-prone response variable, the amount of measurement error in the covariates of the predictive model, the size of the validation dataset and the number of imputations.

CHAPTER 3. STUDY METHODS

This chapter describes the MI methods using Frequentist and Bayesian logistic regression model as predictive model for a main study/validation study design. Using the MI method, K >1 complete datasets are obtained by replacing the missing true values with K simulated values obtained from independent draws from a predictive model. The estimates of the interested parameters from K complete datasets are combined by averaging. The variability of the estimates includes both within-imputation and between-imputation variance. The following sections describe the characteristics of the validation and main datasets, measurement error/misclassification models, predictive models, inference of MI and the underlying assumptions of the models.

3.1 Validation and Main Dataset Characteristics

Let n denote the number of subjects in the validation dataset and N denote the number of subjects in the entire dataset. The number of subjects in the main dataset, in which the information on true disease status is missing, is $N - n$. The error-free measure of disease status (i.e., presence/absence) is denoted by Y. Note that Y is not observed in the main dataset but it is observed in the validation dataset. The observed, error-prone measure of disease status, denoted by U, is observed in both the validation and main datasets. Specificity and sensitivity of the observed disease status are defined as, $SP = P(U = 0|Y = 0)$ and $SN = P(U = 1|Y = 1)$, respectively. In this study, we consider the case in which the observed response is under-reported (i.e., sensitivity < 1.0) but there are no false positives (i.e., specificity is equal to 1.0).

The disease model considered in this study is one in which two covariates are associated with the probability of disease presence in the validation and main datasets[42,43,48]. These are a continuous covariate denoted by X_1 and a categorical (i.e., binary) covariate denoted by X_2. The covariate X_1 is assumed to follow a normal distribution with parameters μ_0 and σ_0^2 denoting the mean and variance, respectively. The covariate X_2 is assumed to follow a binomial distribution with parameters B=1 and p_B denoting the number of trials (equals to 1) and the probability of success, respectively. We assume that X_1 and X_2 are independent. Formally, the disease model is given by

$$P\left(Y_j = 1 \middle| X_{1j}, X_{2j}; \boldsymbol{\beta}\right) = f_{Y|X_1,X_2}(x_{1j}, x_{2j}; \boldsymbol{\beta}) = \frac{e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j})}}{1+e^{(\beta_0+\beta_1 x_{1j}+\beta_2 x_{2j})}}, \qquad (3\text{-}1)$$

for $j = 1, \ldots, N$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$ is the vector of regression parameters and $^{\mathrm{T}}$ denotes the transpose operator.

Let W_1 and W_2 denote the error-prone measures of the true covariates X_1 and X_2, respectively. As well, an indicator variable, V_j, indicates whether the *j*th subject is in the validation dataset or main dataset. If $V_j = 1$ the j^{th} subject is in the validation dataset and if $V_j = 0$ the j^{th} subject is in the main dataset. To summarize, the validation dataset contains Y, U, X_1, X_2, W_1 and W_2, while the main dataset only contains U, W_1 and W_2 (see Figure 3-1). Missing values are denoted by '.'.

	$\boldsymbol{V}$	$\boldsymbol{Y}$	$\boldsymbol{U}$	$\boldsymbol{X_1}$	$\boldsymbol{W_1}$	$\boldsymbol{X_2}$	$\boldsymbol{W_2}$
Validation Dataset	1	1	1	12	12	1	0
	1	1	0	16	14	0	1
	1	0	0	9	10	1	1
	1	1	0	5	6	0	0
	1	0	0	20	21	1	0
Main Dataset	0	.	1	.	23	.	1
	0	.	0	.	6	.	0
	0	.	0	.	15	.	1
	0	.	0	.	17	.	1
	0	.	1	.	13	.	0
	…	…	…	…	…	…	…

Figure 3-1 An illustration of the validation dataset/main dataset design

3.2 Measurement Error and Misclassification in Disease Predictors

Measurement error under a continuous covariate X_1 can be considered under both the Classical and Berkson models[46,79]. As defined previously, the Classical model is $W_1 = X_1 + \varepsilon_1$, where ε_1 is uncorrelated with X_1 that is $E(\varepsilon_1|X_1) = 0$. In this model W_1 is an unbiased measurement of X_1. The structure of ε_1 is homoscedastic. The Berkson model is defined as $X_1 = W_1 + \varepsilon_1$, where ε_1 is uncorrelated with W_1, that is $E(\varepsilon_1|W_1) = 0$.

Let $f_{X_1}(x_1)$ denote the marginal distribution of X_1. The Classical model can be incorporated into a model to predict X_1 via Bayes theorem,

$$f_{X_1|W_1}(x_1|w_1) = \frac{f_{W_1|X_1}(w_1|x_1)f_{X_1}(x_1)}{\int f_{W_1|X_1}(w_1|x_1)f_{X_1}(x_1)dx_1} \qquad (3\text{-}2)$$

where $f_{W_1|X_1}(w_1|x_1)$ is the density of W_1 given X_1. Having $f_{X_1}(x_1)$ the density of X_1, equation (3-2) can transform the Classical model into a predictive model. Thus the predictive model for the continuous covariate X_1 is

$$f_{X_1|W_1}(x_{1i}|w_{1i}, \boldsymbol{\varphi}) = \varphi_0 + \varphi_1 w_{1i} + \varepsilon_1, \qquad (3\text{-}3)$$

where $i = n+1, \ldots, N$, $\boldsymbol{\varphi} = (\varphi_0, \varphi_1)^{\mathrm{T}}$ and $\varepsilon_1 \sim N(0, \sigma^2)$ which is independent of X_1 and W_1. The function $f_{X_1|W_1}(x_{1i}|w_{1i}, \boldsymbol{\varphi})$ is the calibration function of true continuous covariate conditional on the measured continuous covariate and can be used to predict the true value of the continuous covariate. The predictive model for the binary covariate is,

$$f_{X_2|W_2}(x_{2i}|w_{2i}, \boldsymbol{\delta}) = \frac{e^{(\delta_0+\delta_1 w_{2i})}}{1+e^{(\delta_0+\delta_1 w_{2i})}} \qquad (3\text{-}4)$$

where $\boldsymbol{\delta} = (\delta_0, \delta_1)^{\mathrm{T}}$ is the vector of parameters reflecting the association between X_2 and W_2. The function $f_{X_2|W_2}(x_{2i}|w_{2i}, \boldsymbol{\delta})$ is the calibration function of true binary covariate conditional on the measured binary covariate and can be used to predict the true value of the binary covariate.

3.3 Development of the Predictive Model

In this study, a logistic regression model is used to build a predictive model for disease status in the validation dataset. Two models were considered:

3.3.1 Scenario 1: Misclassification of Observed Disease Status is Independent of Disease Predictors

The first scenario is one in which misclassification of the observed disease status is assumed to be independent of the disease predictors in the validation dataset. In this case, true disease status is predicted from observed disease status only. Thus, the predictive model is

$$logit\ P(Y_l|U_l) = \alpha_0 + \alpha_1 U_l, \qquad (3\text{-}5)$$

where $l = 1, \dots n$. This predictive model is fitted to the validation dataset to estimate the parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^{\mathrm{T}}$. The predictive model for the main dataset is

$$logit\ P(Y_i|U_i) = \hat{\alpha}_0 + \hat{\alpha}_1 U_i \qquad (3\text{-}6)$$

where $i = n + 1, \dots, N$. Using equation (3-6), the probability of being a disease case is estimated. This probability becomes the parameter for an observation sampled from the Bernoulli distribution. Specifically, Bernoulli ($P(Y_i = 1|U_i)$) generates a binary response for the ith subject in the main dataset.

To estimate the parameters of this logistic regression model, the most widely used method is the iterative Newton-Raphson algorithm,

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U})^{-1} g(\boldsymbol{\alpha}) \qquad (3\text{-}7)$$

where $\boldsymbol{\alpha}^{(t)}$ is the parameter vector for the t^{th} iteration, $g(\boldsymbol{\alpha}) = \boldsymbol{U}^{\mathrm{T}}(\boldsymbol{Y} - \widehat{\boldsymbol{\theta}})$, and $\boldsymbol{\Lambda} = \mathrm{diag}(\widehat{\boldsymbol{\theta}}(1 - \widehat{\boldsymbol{\theta}}))$ with $\widehat{\boldsymbol{\theta}} = \frac{1}{1+e^{-(\alpha_0+\alpha_1 U)}}$.

When the sample size or the total Fisher information is small, there is the potential for bias in the maximum likelihood estimates. The amount of bias in the estimated parameters for linear logistic models, as proposed by Cordeiro and McCullagh[80], is

$$\boldsymbol{b} = (\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\eta} \qquad (3\text{-}8)$$

where $\boldsymbol{\eta} = \mathrm{diag}[\boldsymbol{U}(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{U})^{-1}\boldsymbol{U}^{\mathrm{T}}](\widehat{\boldsymbol{\theta}} - 0.5)$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\widehat{\boldsymbol{\theta}}(1 - \widehat{\boldsymbol{\theta}}))$. Therefore, a bias-corrected estimate, was proposed by the authors, and is given by $\widehat{\boldsymbol{\alpha}}_C = \widehat{\boldsymbol{\alpha}} - \boldsymbol{b}$.

In studies that validate AHDs, the validation dataset is often small[81,82], therefore we compared both the bias-corrected estimates and uncorrected estimates for the Frequentist approach. Using the bias-reduced estimates $\widehat{\boldsymbol{\alpha}}_C$ and covariance matrix C($\widehat{\boldsymbol{\alpha}}_C$), multiple values of the parameters $\widetilde{\boldsymbol{\alpha}}^{(k)}$ $(k = 1, \dots, K)$ are drawn from its asymptotic normal posterior distribution $\widetilde{\boldsymbol{\alpha}}^{(k)}$~N[$\widehat{\boldsymbol{\alpha}}_C$, C($\widehat{\boldsymbol{\alpha}}_C$)]. The corresponding values of the probability $P(Y_i^{(k)}|U, \widetilde{\boldsymbol{\alpha}}^{(k)})$ are obtained from equation 6 and multiple imputed values of the true response are drawn from the Bernoulli distribution $Y_{\mathrm{i}}^{(k)}$~Benoulli($P\left(Y_i^{(k)}\middle|U, \widetilde{\boldsymbol{\alpha}}^{(k)}\right)$).

For the Bayesian approach, we used a previously proposed logistic regression model[83]. Specifically, the authors used a prior in the conjugate form as the likelihood function, so the posterior had the same form as the likelihood as well. Then the value of the interested parameter that maximized the posterior was the maximum posterior estimator. Generally, the Markov chain Monte Carlo (MCMC) algorithms such as the Metropolis-Hastings and the Gibbs sampler can be implemented to sample the parameters of interest from their posterior distributions[84]. The predictive model in equation (3-6) was adopted. The parameters $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)^{\mathrm{T}}$ were treated as random variables. Non-informative priors were chosen; these follow a normal distribution with a mean of zero and large variance. Multiple values of the parameters $\hat{\alpha}_k$ $(k = 1, \dots, K)$ were randomly drawn from the posterior distribution and used to compute the disease probability, that is $P\left(Y_i^{(k)}\middle|U, \widehat{\boldsymbol{\alpha}}^{(k)}\right)$. Subsequently, multiple imputed values of the true response were drawn from the Bernoulli distribution $Y_{\mathrm{i}}^{(k)}$~Benoulli($P(Y_i^{(k)}|U, \widehat{\boldsymbol{\alpha}}^{(k)})$).

3.3.2 Scenario 2: Misclassification of Observed Disease Status is Dependent on Disease Predictors

For this scenario, the misclassification of the observed disease status is conditional on one or more disease predictor variables. Under the Frequentist approach, a predictive logistic model was developed in the validation dataset using two covariates, one of which was binary and one of which was continuous, and the observed disease status. From equation (3-3) and equation (3-4), the bias-reduced parameters ($\widehat{\boldsymbol{\varphi}}_C$ and $\widehat{\boldsymbol{\delta}}_C$) and the covariance matrixes ($C(\widehat{\boldsymbol{\varphi}}_C)$ and $C(\widehat{\boldsymbol{\delta}}_C)$) were estimated. Multiple values of the parameters of each model were drawn from the asymptotic normal distribution N[$\widehat{\boldsymbol{\varphi}}_C$, C($\widehat{\boldsymbol{\delta}}_C$)] and N[$\widehat{\boldsymbol{\delta}}_C$, C($\widehat{\boldsymbol{\delta}}_C$)]. The multiple predicted values of the true covariates were obtained using the equations $\hat{X}_{1i}^{(k)} = \hat{\varphi}_0^{(k)} + \hat{\varphi}_1^{(k)} W_{1i}$ and

$\hat{X}_{2i}^{(k)} \sim Bernoulli(P(\hat{X}_{2i}^{(k)}|W_{2i}, \widehat{\boldsymbol{\delta}}^{(k)}))$ with $P\left(\hat{X}_{2i}^{(k)}\middle|W_{2i}, \widehat{\boldsymbol{\delta}}^{(k)}\right) = \frac{e^{(\hat{\delta}_0^{(k)} + \hat{\delta}_1^{(k)} W_{2i})}}{1+e^{(\hat{\delta}_0^{(k)} + \hat{\delta}_1^{(k)} W_{2i})}}$, where $k = 1, \ldots, K$. The predictive model of the true response fitted in the validation dataset to estimate the parameter vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)^{\mathrm{T}}$ was

$$logit\, P(Y_l|\, X_{1l}, X_{2l}) = \gamma_0 + \gamma_1 X_{1l} + \gamma_2 X_{2l} + \gamma_3 U_l. \qquad (3\text{-}9)$$

The model to predict the missing true response in the main dataset is

$$logit\, P\left(Y_i = 1\middle|\, \hat{X}_{1i}, \hat{X}_{2i}\right) = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{X}_{1i} + \hat{\gamma}_2 \hat{X}_{2i} + \hat{\gamma}_3 U_i\,. \qquad (3\text{-}10)$$

From equation (3-9) the bias-reduced estimates were used to produce the estimates $\widehat{\boldsymbol{\gamma}}_C$ and $C(\widehat{\boldsymbol{\gamma}}_C)$. Multiple values of $\widehat{\boldsymbol{\gamma}}_C$ were drawn from the asymptotic normal distribution ~N[$\widehat{\boldsymbol{\gamma}}_C$, C($\widehat{\boldsymbol{\gamma}}_C$)]. To impute Y in the main dataset, we apply $\widehat{\boldsymbol{\gamma}}_C$ and estimated covariates $\hat{X}_1$ and $\hat{X}_2$ to equation (3-10) to obtain $logit\, P\left(Y_i^{(k)} = 1\middle|\, \hat{X}_{1i}, \hat{X}_{2i}\right) = \hat{\gamma}_0 + \hat{\gamma}_1 \hat{X}_{1i} + \hat{\gamma}_2 \hat{X}_{2i} + \hat{\gamma}_3 U_i$. The multiple imputed values of the missing true response in the main dataset Y_{M} were drawn from the Bernoulli distribution, $Y_i^{(k)} \sim$Benoulli $\left(P\left(Y_i^{(k)}\middle|U, \widetilde{\boldsymbol{\alpha}}^{(k)}\right)\right)$.

Three different predictive models were considered. The predictive model defined in equation (3-9) was compared with two other predictive models:

$$logit\, P(Y_l|\, X_{1l}, X_{2l}) = \gamma_0 + \gamma_1 \hat{X}_{1l} + \gamma_2 \hat{X}_{2l} \qquad (3\text{-}11)$$

$$logit\, P(Y_l|\, X_{1l}, X_{2l}) = \gamma_0 + \gamma_1 U_l. \qquad (3\text{-}12)$$

These model comparisons were done because in practice, different variables might be selected as covariates in the predictive model.

3.4 Prevalence and Variance Estimation

Using the imputed disease status values computed in the previous section, the estimated prevalence for the k_{th} dataset is $\hat{p}^{(k)} = \frac{\sum Y_l + \sum Y_i^{(k)}}{N}$. For the MI method, the final estimated prevalence for the entire population is $\hat{p} = \frac{1}{K}\sum_{k=1}^{K} \hat{p}^{(k)}$, which reflects the uncertainty caused by

the missingness of true values. The total variability, T_K, includes within-imputation variance ($\bar{W}_K$) and between-imputation variance (B_K), that is,

$$\bar{W}_K = \frac{1}{K}\sum_{k=1}^{K} W_k \qquad (3\text{-}13)$$

where $W_k = Var\left(\hat{p}^{(k)}\right) = Var\left(\frac{\sum Y_l + \sum Y_i^{(k)}}{N}\right) = \frac{1}{N^2}\sum_{i=n+1}^{N} Var(Y_i^{(k)}) = \frac{1}{N^2}\sum_{i=n+1}^{N} P\left(Y_i^{(k)}\right)[1 - P(Y_i^{(k)})]$,

$$B_K = \frac{1}{K-1}\sum_{k=1}^{K}(\hat{p}^{(k)} - \hat{p})^2 \qquad (3\text{-}14)$$

$$T_K = \bar{W}_K + \frac{K+1}{K} B_K. \qquad (3\text{-}15)$$

With the final estimated prevalence $\hat{p}$ and the total variability T_K, the 95% confidence interval was calculated based on asymptotic normal distribution.

CHAPTER 4. SIMULATION STUDY

A Monte Carlo simulation study was undertaken to investigate the MI methods for measurement error correction. The parameters that were manipulated and investigated in the Monte Carlo study include sensitivity of the observed disease status, amount of error in the covariates, size of the validation dataset, and numbers of imputations. Table 4-1 summarizes the conditions that were investigated for each of these parameters. The mechanism by which observations were missing, including MCAR, MAR and MNAR conditions were also examined.

Table 4-1: Parameters of the simulation study

Sensitivity of observed disease status	Measurement error in W_1	Misclassification in W_2	Size of validation dataset	Number of imputations
0.60	Var(ε_1)=1 or 2	SN=SP=0.7 or SN=SP=0.9	0.05, 0.20, 0.35	3, 5, 10, 15, 20, 40
0.75	Var(ε_1)=1 or 2	SN=SP=0.7 or SN=SP=0.9		
0.90	Var(ε_1)=1 or 2	SN=SP=0.7 or SN=SP=0.9		

Note: SN denotes sensitivity and SP denotes specificity

The Monte Carlo study was conducted using SAS/IML (Interactive Matrix Language) software version 9.2. The simulation program is included in Appendix D. The RANGEN procedure was used to generate random numbers from the specified distributions. A total of M=500 replications was conducted for each combination of conditions. A population size of N=10,000 was used to implement the models.

The data were generated using the disease model defined in the previous chapter. The continuous covariate, X_1 was generated from the standard normal distribution (i.e., $N(0,1)$) and the binary covariate, X_2 was generated from the binomial distribution (i.e., BIN(N, p_B)). These two covariates were generated independently. The true response was generated based on the disease model (equation 1) using specified values of the regression coefficients β_0, β_1 and β_2. Then the Bernoulli random variable with probability P was generated as 1 or 0, where 1 denotes disease presence and 0 denotes disease absence.

The coefficients β_1 and β_2 denote the strength of association between X_1, X_2 and the true response Y. The values of these coefficients were selected based on the odds ratio, which is obtained by exponentiation of the regression coefficient. The choice of $\beta_1 = 0.269$ corresponds to an odds ratio of 1.3 while $\beta_1 = 0.539$ corresponds to an odds ratio of 2.0 between the 10th and the 90th percentile of the standard normal covariate. The choice of $\beta_2 = -0.693$ corresponds to an odds ratio of 0.5 while $\beta_2 = 1.386$ produces an odds ratio of 2.0 between 0 and 1 of the Bernoulli distributed covariate. These odds ratios are realistic for epidemiological studies about chronic disease[85-87]. We specified $\beta_0 = -2.785$, $\beta_1 = 0.539$ and $\beta_2 = -0.693$ to generate the data has prevalence of the true response is 0.05, because of the low prevalence of chronic disease in the population.

The methods to generate the observed response are different for the two scenarios of the sensitivity of the observed response. For Scenario 1, under-reporting is independent of the covariates, that is, when $Y = 1$ the probability of $U = 1$ is randomly generated from the Bernoulli distribution with π_1, the sensitivity of the response variable, as the parameter of this distribution. Values for π_1 from 0.6 to 0.9 in increments of 0.15 were considered. For Scenario 2, under-reporting depends on the covariates, that is when $Y = 1$ the probability of $U = 1$ was generated based on the association with the model covariates. By specifying the coefficients of the generation model for observed disease status, the observed response also had sensitivity from 0.6 to 0.9 in increments of 0.15.

For Scenario 2, W_1 was generated by adding the additive variance as Classical error model[88]. Specifically, the additive variance was generated from a normal distribution with variance of 1 (i.e., N(0,1)) or variance of 2 (i.e., N(0,2)). And the measurement error of W_1 was assumed non-differential. The observed binary covariate was generated by manipulating values of sensitivity and specificity for two conditions. In Condition 1, the sensitivity and specificity of the mismeasured binary covariate were equal to 0.7. In Condition 2, both the sensitivity and specificity were equal to 0.9. Specifying the coefficients of the true disease predictors generated the observed disease status with sensitivity from 0.6 to 0.9 in increments of 0.15, because the misclassification of the observed disease status is dependent on the true values of the disease predictors.

The mechanisms of the missingness of true response were also considered in the study design. If *V*, the indicator of the validation dataset, is a random variable then the missing mechanism is MCAR. If *V* is dependent on the observed response not on the missing true response then the missing mechanism is MAR. And if *V* is dependent on the missing true response then the missing mechanism is MNAR.

In terms of the size of the validation dataset, we considered cases where it was a fixed proportion of the entire dataset: 0.05*N*, 0.20*N* and 0.*35N*. The number of imputations took on values of 3, 5, 10, 15, 20 and 40[15], in keeping with previous research.

As for the Bayesian logistic regression, the MCMC procedure, which is available in SAS 9.2 was adopted. It uses a random walk Metropolis algorithm to simulate samples from the specified logistic regression model[89]. A total of 21,000 iterations were conducted in the simulation loop of MCMC procedure, and the first 1000 iterations were burn-in iterations. Thus, the last 20,000 iterations, with the thinning rate of 100, were saved as posterior samples. That means every 100th simulation sample of the 20,000 iterations was kept and the rest were discarded, which results in 200 posterior samples. The posterior statistics and diagnostics such as MCSE/SD (the Monte Carlo standard errors of each parameter relative to the posterior standard deviations) and autocorrelation were calculated using the thinned posterior samples.

Bias, relative bias, RMSE and coverage were used to evaluate the performance of the models,

$$Bias = \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{K}\sum_{k=1}^{K}(\hat{p}_{mk} - p)), \quad (4\text{-}1)$$

$$Relative\ Bias = \frac{bias}{true\ prevalence}, \quad (4\text{-}2)$$

$$MSE = variance + bias^2, \quad (4\text{-}3)$$

where M is the number of replications, K is the number of imputations, the variance is the total variability as defined in equation (3-15), and $\hat{p}_{mk}$ is the estimate of prevalence in the mth replicated dataset and kth imputation. To put the measure on the same scale as the prevalence, we report the square root of the MSE (RMSE). Coverage was calculated as the proportion of the simulations in which the 95% confidence intervals covered the true parameter value.

CHAPTER 5. RESULTS

5.1 Scenario 1: Misclassification of Observed Disease Status is Independent of Disease Predictors

Table 5-1: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.05N$

Sensitivity of observed disease status	Missing mechanism	Bayesian	Frequentist	Frequentist (Bias Corrected)
			Relative Bias (%)	
0.60	MCAR	-37.13	-26.14	-0.32
	MAR	-37.69	-28.34	-6.61
	MNAR	-39.69	-23.57	-32.03
0.75	MCAR	-23.53	-31.36	-0.69
	MAR	-23.75	-33.66	-6.86
	MNAR	-25.01	62.97	-23.16
0.90	MCAR	-9.53	37.61	-0.31
	MAR	-9.81	27.18	-6.99
	MNAR	-10.33	310.49	-13.05
			RMSE	
0.60	MCAR	0.0185	0.0145	0.0031
	MAR	0.0188	0.0155	0.0048
	MNAR	0.0197	0.0323	0.0162
0.75	MCAR	0.0117	0.0198	0.0027
	MAR	0.0118	0.0192	0.0048
	MNAR	0.0125	0.0678	0.0120
0.90	MCAR	0.0047	0.0497	0.0022
	MAR	0.0049	0.0466	0.0048
	MNAR	0.0052	0.1903	0.0072
			Coverage	
0.60	MCAR	0.00	0.00	0.53
	MAR	0.00	0.00	0.50
	MNAR	0.00	0.00	0.05
0.75	MCAR	0.00	0.41	0.53
	MAR	0.00	0.00	0.50
	MNAR	0.00	0.00	0.10
0.90	MCAR	0.00	0.49	0.60
	MAR	0.01	0.00	0.50
	MNAR	0.00	0.01	0.26

Note: RMSE = root mean squared error; MCAR is missing completely at random; MAR is missing at random; MNAR is missing not at random

Results for Scenario 1 are reported first. The simulation results are described for each of the measures of relative bias, RMSE, and coverage probability. Tables 5-1 to 5-3 present the results for the three different sizes of the validation dataset.

When the size of the validation dataset was 0.05 *N* (Table 5-1), the Bayesian MI model resulted in relative bias and RMSE values for the MCAR condition that were smaller than for the MAR and MNAR conditions. For the Frequentist MI model without bias correction, the relative bias and RMSE were the largest when the sensitivity was 0.90 under the MNAR condition. The Frequentist MI model with bias correction had substantially smaller relative bias for the MAR condition (70.02%) than for the MNAR condition. When the sensitivity of the observed disease status was 0.75 or 0.90, the Bayesian MI model had smaller relative bias than the Frequentist model without bias correction. However, the Frequentist model without bias correction had better coverage when the missingness mechanism was MCAR and the sensitivity was 0.75 or 0.90. In this situation, the Frequentist model with bias correction had the smallest relative bias; absolute values were less than 1.00% under the MCAR condition and less than 7.00% under the MAR condition. The 95% confidence intervals for the Frequentist model with bias correction had coverage probability greater than 0.50 if the MAR assumption was not violated. As for the effect of the sensitivity of the observed disease status, the results revealed that different methods behaved differently as the sensitivity increased. For the Bayesian model, the relative bias and RMSE became smaller when the sensitivity increased. For the Frequentist model with bias correction, only under the MNAR condition did the relative bias and RMSE decrease with increasing sensitivity.

Table 5-2: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.20N$

Sensitivity of observed disease status	Missing mechanism	Bayesian	Frequentist	Frequentist (Bias Corrected)
			Relative Bias (%)	
0.60	MCAR	-31.32	-23.36	-0.33
	MAR	-31.18	-27.80	-2.43
	MNAR	-37.26	-53.61	-28.73
0.75	MCAR	-19.66	-29.02	-0.06
	MAR	-19.67	-34.72	-2.49
	MNAR	-23.36	-50.76	-18.97
0.90	MCAR	-7.98	-35.12	-0.25
	MAR	-7.99	-41.68	-2.55
	MNAR	-9.42	-24.07	-9.07
			RMSE	
0.60	MCAR	0.0156	0.0126	0.0018
	MAR	0.0155	0.0150	0.0023
	MNAR	0.0185	0.0273	0.0143
0.75	MCAR	0.0098	0.0156	0.0015
	MAR	0.0098	0.0186	0.0021
	MNAR	0.0116	0.0263	0.0095
0.90	MCAR	0.0040	0.0189	0.0011
	MAR	0.0040	0.0223	0.0019
	MNAR	0.0047	0.0287	0.0047
			Coverage	
0.60	MCAR	0.00	0.00	0.72
	MAR	0.00	0.00	0.71
	MNAR	0.00	0.00	0.00
0.75	MCAR	0.00	0.50	0.73
	MAR	0.00	0.00	0.69
	MNAR	0.00	0.00	0.01
0.90	MCAR	0.00	0.51	0.76
	MAR	0.00	0.00	0.66
	MNAR	0.00	0.00	0.06

Note: RMSE = root mean squared error; MCAR is missing completely at random; MAR is missing at random; MNAR is missing not at random

When the size of the validation dataset was $0.20N$ (Table 5-2), the results followed similar trend to those when the size of the validation dataset was $0.05N$. However, the RMSE values for all MI models were smaller than when the size of the validation dataset was $0.05N$. And in this situation, measures for each model under MCAR and MAR conditions were similar, which were

quite different from the measures for the model under MNAR condition. The Bayesian MI model had smaller relative bias than the Frequentist MI model except when the sensitivity was 0.60 and under MCAR and MAR. Under MCAR and when the sensitivity was 0.75 and 0.90, the Frequentist model without bias correction had coverage 0.50 and 0.51 much better than other conditions for this model. The Frequentist model with bias correction had the best performance in terms of coverage probability and absolute values of relative bias, as long as the missingness mechanism was not MNAR. When the sensitivity increased, the Bayesian MI model had decreasing relative bias and RMSE, but the coverage was unchanged. With the sensitivity increasing, the relative bias and RMSE of the Frequentist MI model increased under MCAR and MAR conditions, and the relative bias decreased under MNAR condition. Under MCAR and MAR conditions, the Frequentist MI model with bias correction had no trend with the sensitivity increasing. Under MNAR, the Frequentist MI model had smaller (33.97%) relative bias when the sensitivity was 0.75 than the relative bias when the sensitivity was 0.60, and had smaller (52.19%) relative bias when the sensitivity was 0.90 than the relative bias when the sensitivity was 0.75.

Table 5-3: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is 0.35N

Sensitivity of observed disease status	Missing mechanism	Bayesian	Frequentist	Frequentist (Bias Corrected)
			Relative Bias (%)	
0.60	MCAR	-25.45	-19.17	-0.14
	MAR	-25.28	-25.94	-1.47
	MNAR	-35.19	-52.58	-27.89
0.75	MCAR	-16.02	-24.02	-0.10
	MAR	-15.82	-32.30	-1.45
	MNAR	-21.99	-48.96	-17.96
0.90	MCAR	-6.50	-28.42	-0.17
	MAR	-6.41	-39.00	-1.45
	MNAR	-8.84	-44.41	-8.06
			RMSE	
0.60	MCAR	0.0127	0.0104	0.0016
	MAR	0.0126	0.0140	0.0019
	MNAR	0.0175	0.0266	0.0139
0.75	MCAR	0.0080	0.0129	0.0013
	MAR	0.0079	0.0173	0.0016
	MNAR	0.0109	0.0252	0.0090
0.90	MCAR	0.0032	0.0152	0.0009
	MAR	0.0032	0.0208	0.0013
	MNAR	0.0044	0.0237	0.0041
			Coverage	
0.60	MCAR	0.00	0.49	0.87
	MAR	0.00	0.00	0.84
	MNAR	0.00	0.00	0.00
0.75	MCAR	0.00	0.50	0.87
	MAR	0.00	0.00	0.84
	MNAR	0.00	0.00	0.00
0.90	MCAR	0.00	0.53	0.87
	MAR	0.00	0.00	0.80
	MNAR	0.00	0.00	0.02

Note: RMSE = root mean squared error; MCAR is missing completely at random; MAR is missing at random; MNAR is missing not at random

Similarly, as Table 5-3 reveals when the size of the validation dataset was 0.35N, the Frequentist MI model with bias correction had the smallest relative bias and RMSE even when the missingness mechanism was MNAR. When the missingness mechanism was MCAR and MAR the coverage of the Frequentist MI model with bias correction was greater than 0.80.

Under MCAR, the Frequentist MI model without bias correction had coverage of 0.49, 0.50 and 0.53 when the sensitivity was 0.60, 0.75 and 0.90, respectively. Under MAR and MNAR, the Frequentist MI model without bias correction had low coverage. Other than when the sensitivity was 0.60 and under MCAR, the Bayesian MI model had smaller relative bias than the Frequentist MI model without bias correction. The relative bias and RMSE of the Bayesian MI model decreased with increasing sensitivity of the observed disease status under all of the three missingness mechanisms. Under the MNAR condition, the Frequentist model with and without bias correction had smaller relative bias and RMSE when sensitivity was larger.

Table 5-4: Relative bias, RMSE and 95% confidence interval coverage with the size of the validation dataset increasing

Missing mechanism	Size of the validation dataset	Bayesian	Frequentist	Frequentist (Bias Corrected)
			Relative Bias (%)	
MCAR	0.05*N*	-23.40	-6.63	-0.44
	0.20*N*	-19.65	-29.17	-0.22
	0.35*N*	-15.99	-23.87	-0.14
MAR	0.05*N*	-23.75	-11.61	-6.82
	0.20*N*	-19.61	-34.73	-2.59
	0.35*N*	-15.84	-32.41	-1.46
MNAR	0.05*N*	-25.01	116.63	-22.75
	0.20*N*	-23.35	-42.81	-18.92
	0.35*N*	-22.01	-48.65	-17.97
			RMSE	
MCAR	0.05*N*	0.0116	0.0280	0.0027
	0.20*N*	0.0098	0.0157	0.0015
	0.35*N*	0.0079	0.0128	0.0013
MAR	0.05*N*	0.0118	0.0271	0.0048
	0.20*N*	0.0097	0.0186	0.0021
	0.35*N*	0.0079	0.0174	0.0016
MNAR	0.05*N*	0.0124	0.0968	0.0118
	0.20*N*	0.0116	0.0274	0.0095
	0.35*N*	0.0109	0.0252	0.0090
			Coverage	
MCAR	0.05*N*	0.00	0.30	0.55
	0.20*N*	0.00	0.34	0.74
	0.35*N*	0.00	0.51	0.87
MAR	0.05*N*	0.01	0.00	0.50
	0.20*N*	0.00	0.00	0.69
	0.35*N*	0.00	0.00	0.83
MNAR	0.05*N*	0.00	0.00	0.13
	0.20*N*	0.00	0.00	0.02
	0.35*N*	0.00	0.00	0.01

Note: MCAR is missing completely at random; MAR is missing at random; MNAR is missing not at random

To assess the effect of the size of the validation dataset, the table above displays the average values of relative bias, RMSE, and coverage over the imputations and sensitivities. The Bayesian MI model did not result in a change in the coverage probability as the size of the validation dataset increased regardless the missingness mechanisms. The relative bias and RMSE of the Bayesian MI model decreased as the size of the validation dataset increased under all three

missingness mechanisms. However, the coverage probability of the Frequentist MI model without bias correction increased as the size of the validation dataset increased under the MCAR condition. When the size of the validation dataset was $0.35N$, the coverage was 0.51, which was much more (70.00%) than the coverage when the size of the validation dataset was $0.05N$. For the Frequentist MI model with bias correction, under the MCAR condition, the average coverage when the size of the validation dataset was $0.35N$ was 17.57% greater than when it was $0.20N$. It was also 34.55% when the size of the validation dataset was $0.20N$ than when it was $0.05N$. Under the MAR condition, the Frequentist MI model with bias correction had better coverage when the size of the validation dataset was larger.

Based on the results in the tables, the Frequentist MI model with bias correction performed better than other methods. So we further explore this model, showing how the prevalence estimates and the standard errors of the prevalence estimates vary with changes in the size of the validation dataset and the number of imputations.

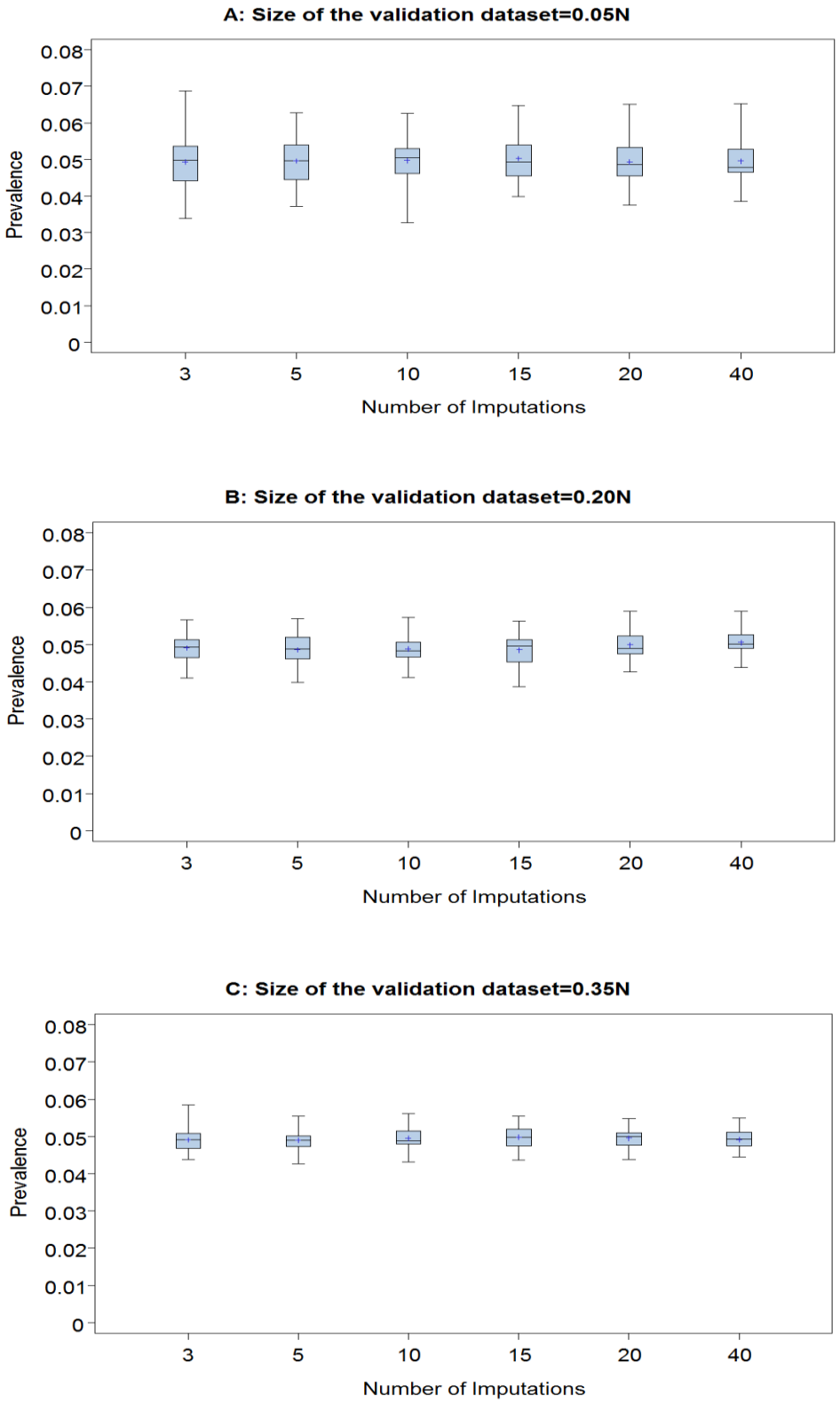

Figure 5-1Prevalence estimates for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.60

Figure 5-1 contains box and whisker plots for the prevalence estimates under the MCAR condition for different sizes of the validation datset when the true population prevalence was 0.05. The estimates were very close to 0.05 regardless of the number of imputations. As panel A reveals, prevalence estimates ranged from 0.03 to 0.07. When the size of the validation dataset was $0.20N$ and $0.35N$, the prevalence estimates ranged from 0.04 to 0.06. The trends of the prevalence estimates for Bayesian MI model and Frequentist MI model were parallel as the trend shown in Figure 5-1 for Frequentist MI model with bias correction.

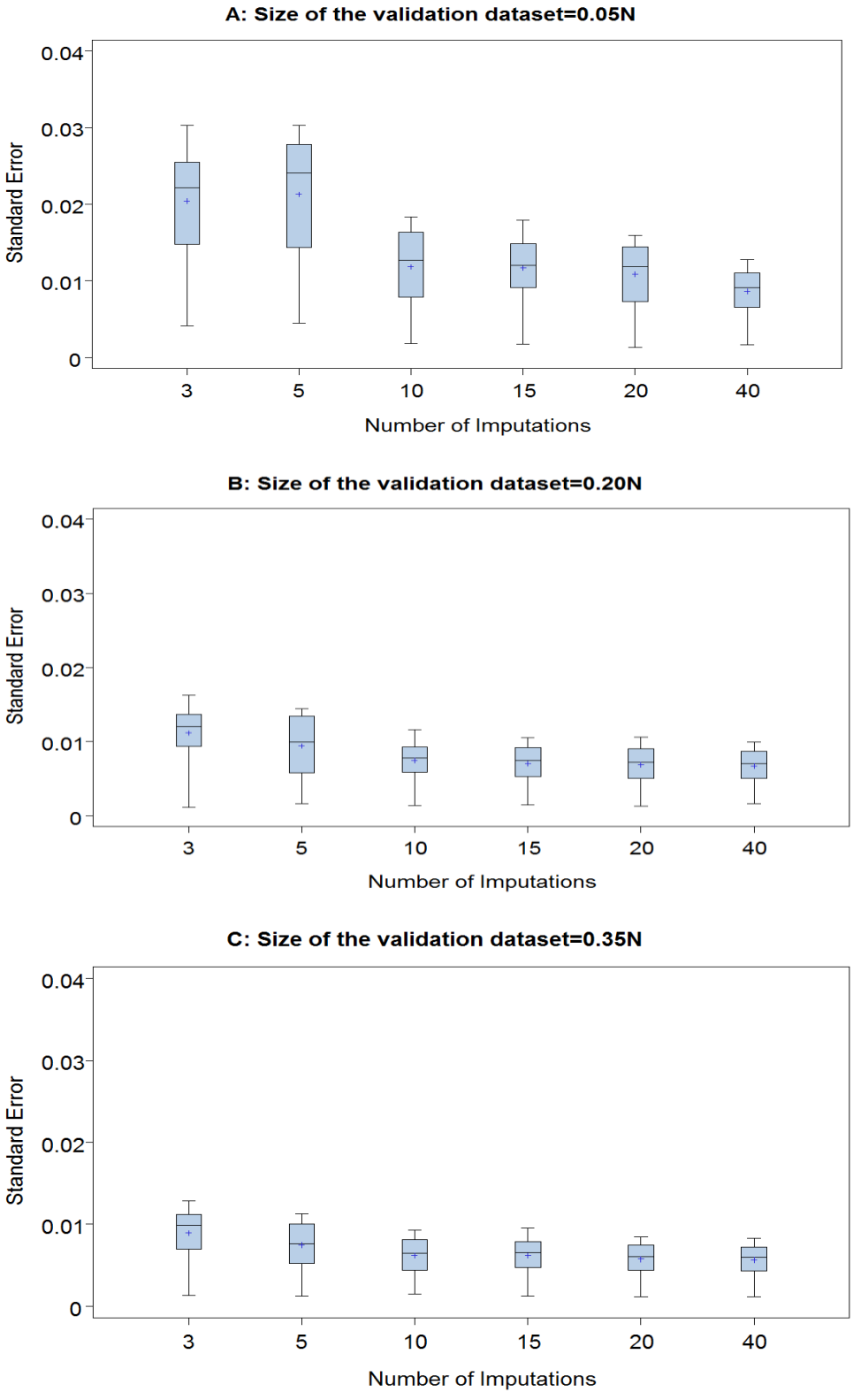

Figure 5-2 Standard error of prevalence estimate for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.60

Figure 5-2 contains box and whisker plots for the standard error values. The standard error decreased as the number of imputations increased. However, when the size of the validation dataset was 0.05N, the standard error decreased significantly (almost 50.00%) when the number of imputations increased from 3 to 10. As the number of imputations increased beyond 10, the change in the standard error was small. When the size of the valdiation dataset was 0.35N, the standard error decreased slightly as the number of imputations increased from 3 to 40. Holding constant the number of imputations, the standard error was about one third smaller when the size of the validation dataset was 0.35N than when it was 0.05N.

For other values of sensitivity, including 0.75 and 0.90, they followed the similar pattern as the sensitivity 0.60 in Appendix A.

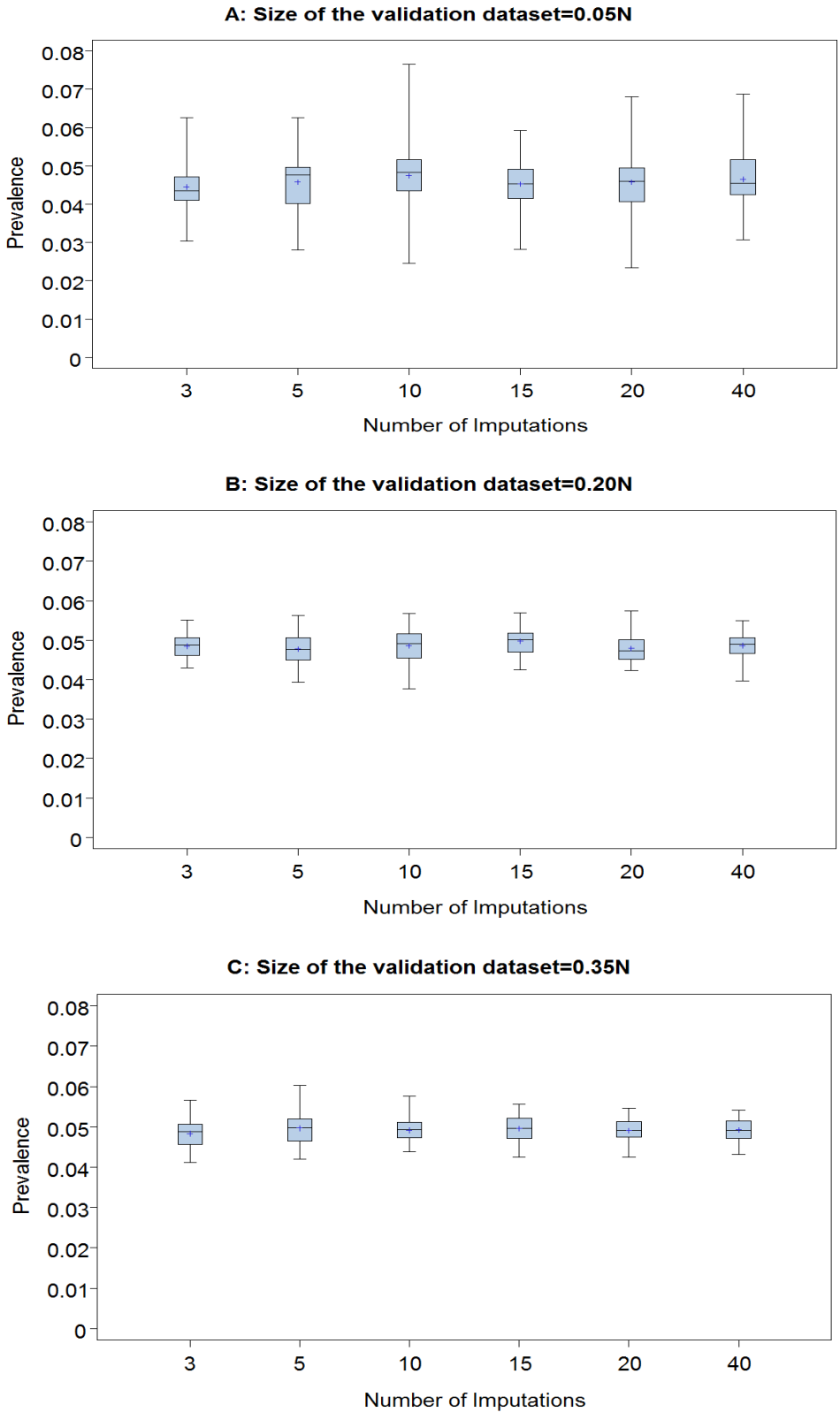

Figure 5-3 Prevalence estimates for the Frequentist MI model with bias correction when the missingness mechanism is MAR and sensitivity is 0.60

Figure 5-3 reveals that under the MAR condition, average prevalence estimates remained stable as the number of imputations increased. When the size of the validation dataset was $0.05N$ the range of values was from 0.02 to 0.08, which was larger than for the MCAR condition (Panel A of Figure 5-1). When the size of the validation dataset was $0.35N$ the range of values for the MAR condition (Panel C of Figure 5-3) was similar to the range for the MCAR condition (Panel C of Figure 5-3). Under MAR, the increase of the size of the validation dataset can decrease the range of the prevalence estimates too. The Bayesian MI model and Frequentist MI model without bias correction had similar pattern as the Frequentist MI model with bias correction.

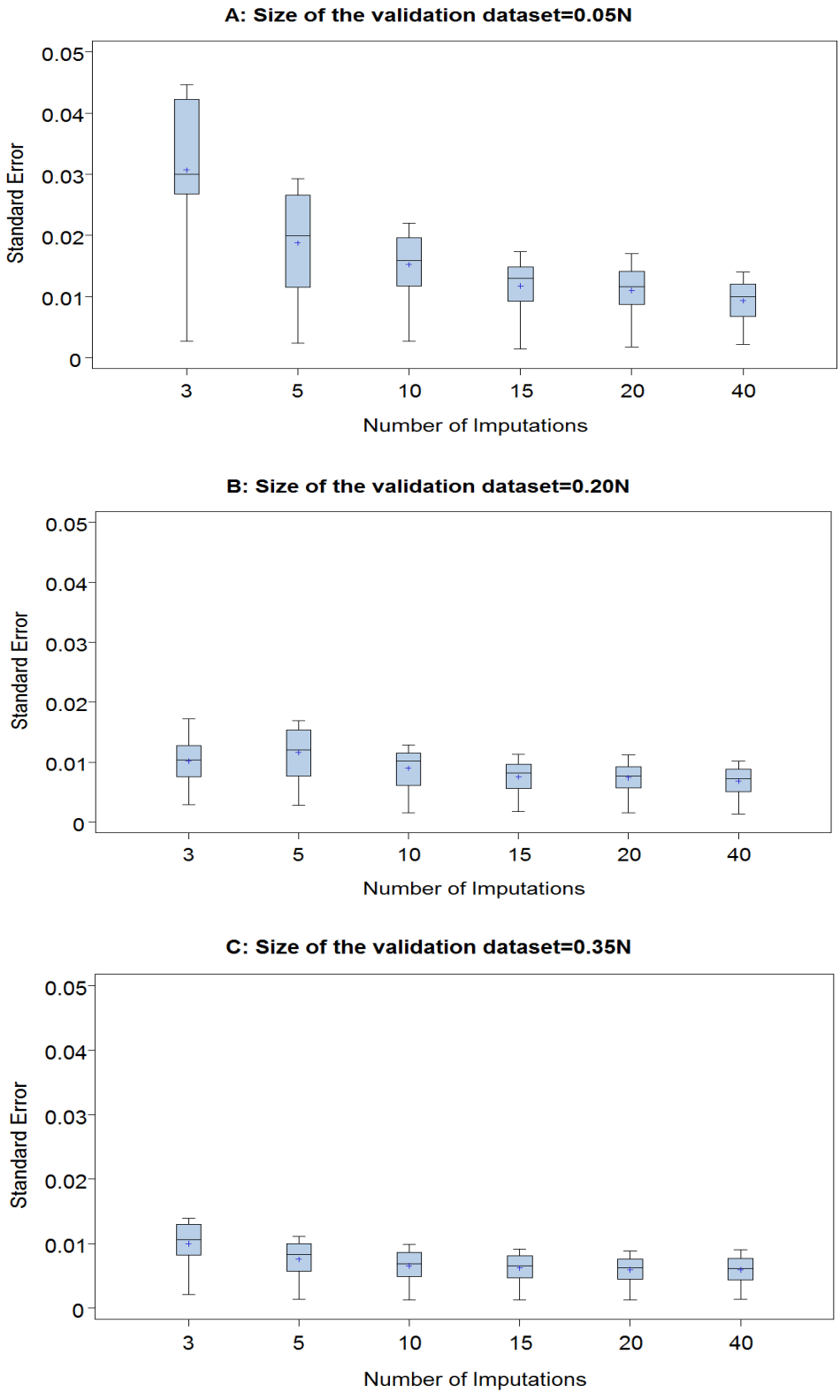

Figure 5-4 Standard errors of prevalence estimates for the Frequentist MI model with bias correction when the missingness mechanism is MAR and sensitivity is 0.60

Figure 5-4 presents the descriptive statistics for the standard error of the prevalence estimates for the MAR condition. When the size of the validation dataset was $0.05N$, the range of values decreased from over 0.04 to slightly more than 0.01 as the number of imputations increased. The standard error was almost 50% smaller for ten imputations than for three imputations. While the size of the validation dataset was $0.20N$ and $0.35N$, there was also a decrease in the size of the standard error, but the magnitude of change was smaller than for the $0.05N$ condition. And the decrease of the standard error when the number of imputations beyond 10 was minor, regardless the size of the validation dataset. When the number of imputations was fixed as 10, the mean of the standard error decreases as the size of the validation increased.

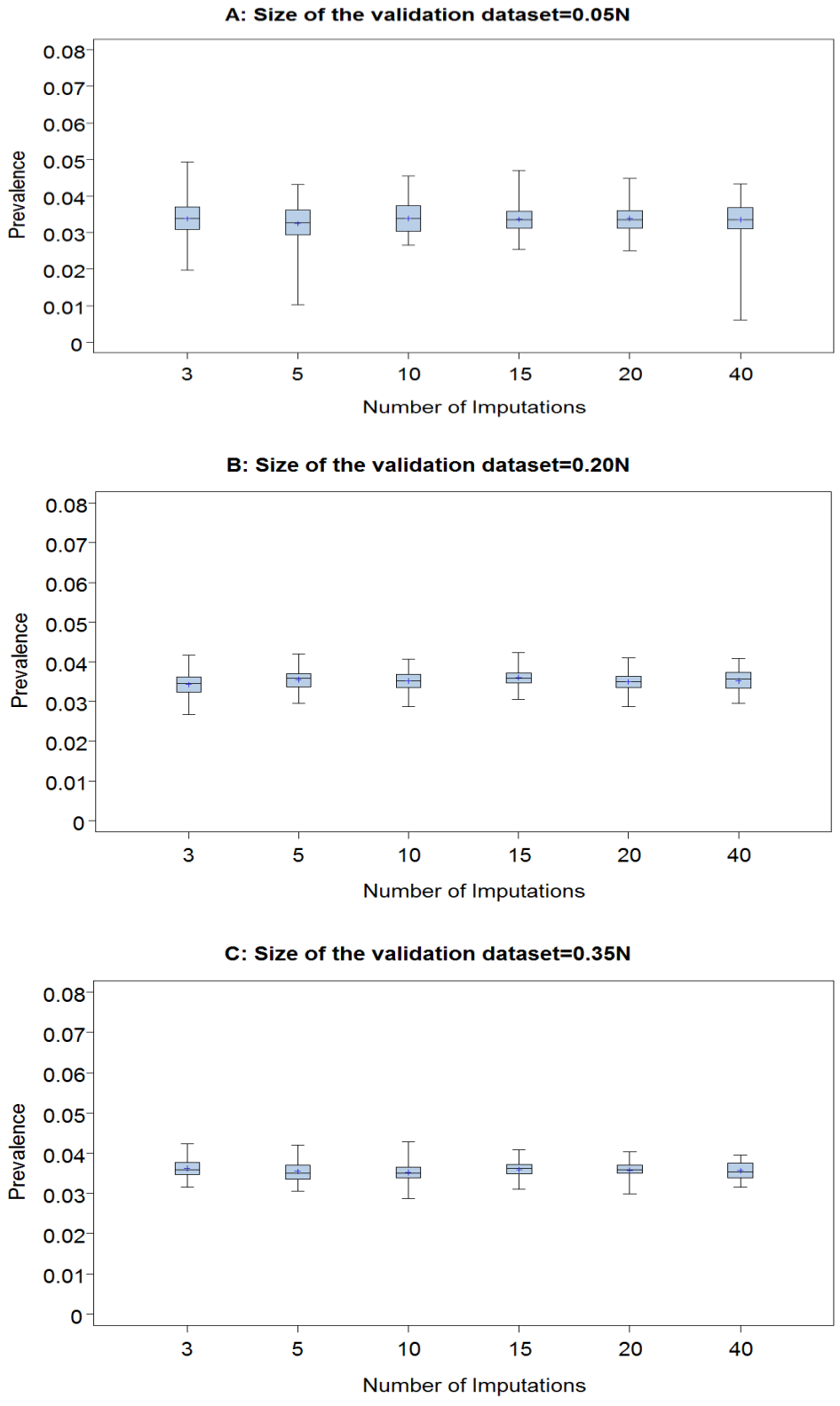

Figure 5-5 Prevalence estimates for Frequentist MI model with bias correction when the missingness mechanism is MNAR and sensitivity is 0.60

From Figure 5-5, we can see that the MNAR mechanism results in a range of prevalence estimates that were also stable when the number of imputations increased. And under MNAR the means of the prevalence estimates were negatively biased for different sizes of the validation dataset. In panel A, the range of prevalence estimates was from less than 0.01 to 0.05. In panel B, the range of prevalence estimate was from a little more than 0.02 to a little over 0.04. In panel C, the range of prevalence estimate was from 0.03 to a little more than 0.04. Although, increase the size of the validation dataset can reduce the range of the prevalence estimates but the means of the prevalence estimates were still biased.

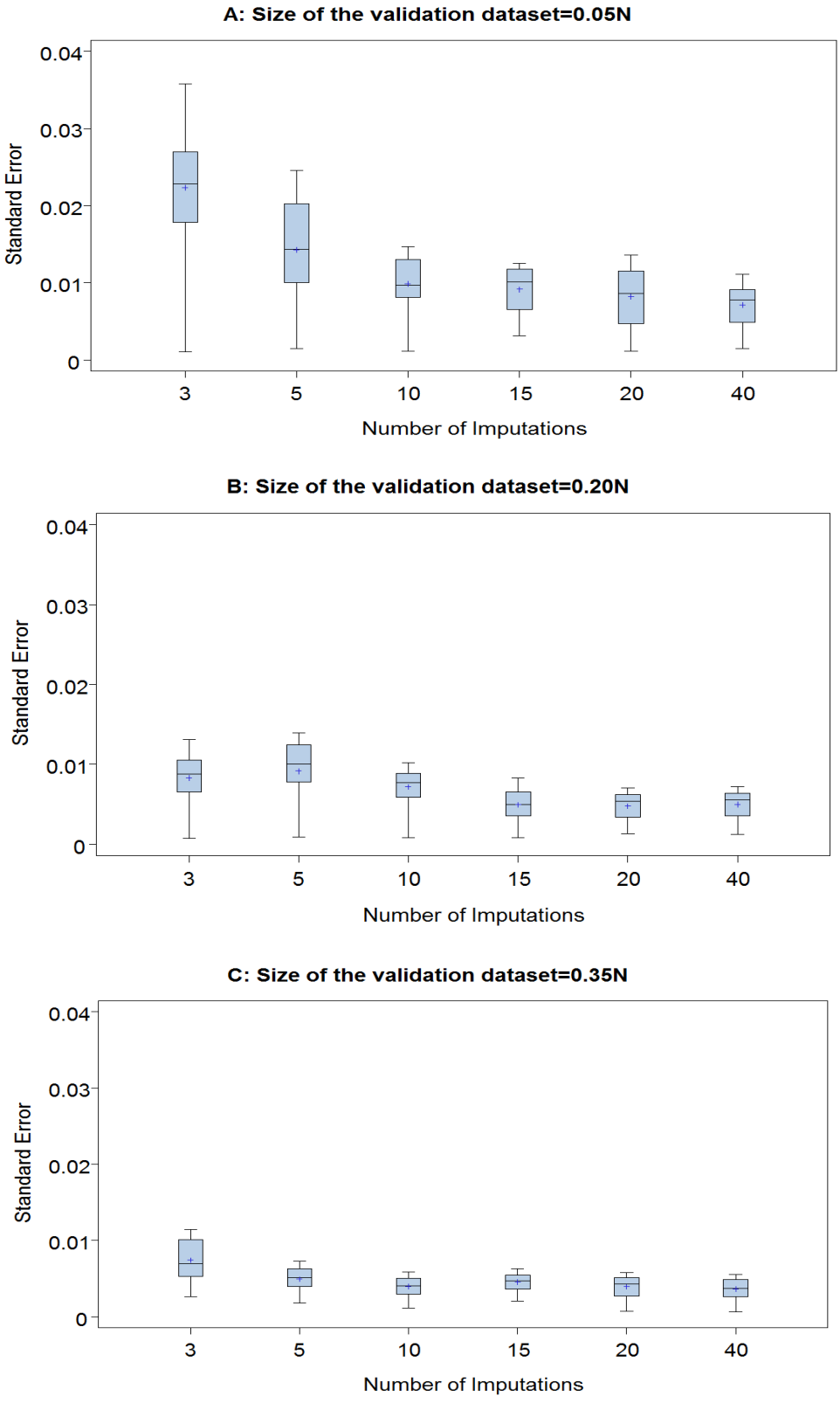

Figure 5-6 Standard error of prevalence estimates for the Frequentist MI method with bias correction when the missingness mechanism is MNAR and sensitivity is 0.60

Figure 5-6 also demonstrates that increasing the number of imputations reduced the range and mean of the standard error of the prevalence estimates under the MNAR condition. When the size of the validation dataset was 0.05*N* the range of standard errors for a small number of imputations was large. When the size of the validation dataset was 0.35*N* the range of standard error was narrower and decreased slightly as the number of imputations increased. However, increasing the number of imputations to more than 10 was not efficient. The trends of the standard errors of prevalence estimates with the number of imputations increasing and the size of the validation dataset increasing under MNAR (Figure 5-6) were similar to those under MCAR (Figure 5-2) and MNAR (Figure 5-4).

5.2 Scenario 2: Misclassification of Observed Disease Status is Dependent on Disease Predictors

In this section, we report the results for the MI models for which when the observed disease status is dependent on the two predictors of true disease status. That means the probability of misclassification of the observed disease status was conditional on the disease predictors. As noted in the previous chapter, Model 1 contained observed disease status and imputed predictors as covariates of the predictive model, while Model 2 has imputed predictors as covariates of the predictive model. Finally, Model 3 has observed disease status the sole covariate of the predictive model. Given the results for Scenario 1, which indicated that the Frequentist MI model with bias correction resulted in less biased estimates of prevalence and generally small values for RMSE, we focused only on the Frequentist MI model with bias correction. However, for comparison, we have included the results for the Frequentist MI model without bias correction in Appendix B.

Table 5-5: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.05N$, Frequentist MI model with bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-0.82	-4.05	-0.47
	MAR	-6.61	-38.99	-5.78
	MNAR	-29.80	-61.02	-31.99
0.75	MCAR	0.32	-4.13	-0.56
	MAR	-6.58	-47.74	-6.83
	MNAR	-29.80	-61.15	-22.97
0.90	MCAR	1.38	-4.06	-0.27
	MAR	-6.67	-55.67	-7.06
	MNAR	-12.46	-61.06	-13.43
			RMSE	
0.60	MCAR	0.0033	0.0048	0.0030
	MAR	0.0050	0.0197	0.0046
	MNAR	0.0152	0.0305	0.0162
0.75	MCAR	0.0029	0.0048	0.0027
	MAR	0.0049	0.0240	0.0048
	MNAR	0.0109	0.0305	0.0118
0.90	MCAR	0.0029	0.0049	0.0023
	MAR	0.0050	0.0279	0.0048
	MNAR	0.0073	0.0305	0.0075
			Coverage	
0.60	MCAR	0.54	0.51	0.52
	MAR	0.50	0.08	0.50
	MNAR	0.07	0.01	0.04
0.75	MCAR	0.56	0.51	0.54
	MAR	0.52	0.03	0.50
	MNAR	0.16	0.01	0.10
0.90	MCAR	0.61	0.51	0.59
	MAR	0.53	0.01	0.50
	MNAR	0.34	0.01	0.24

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table B-1 shows the values of all measures for the Frequentist MI model without bias correction. The values of measures of Model 1 and Model 3 were larger than those of the

Frequentist MI model with bias correction. For the Frequentist MI model without bias correction, the Model 2 had the smallest absolute value of relative bias under MCAR, and the Model 1 had the smallest absolute value of relative bias under MAR when the sensitivity was 0.60 and 0.75. There was no obviously trend of measures with sensitivity increasing for this model too.

Table 5-6: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.20N$, Frequentist MI model with bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-3.14	-24.20	-0.24
	MAR	-5.46	-54.04	-2.34
	MNAR	-29.15	-72.74	-28.63
0.75	MCAR	-2.10	-24.26	-0.21
	MAR	-4.56	-61.06	-2.29
	MNAR	-19.23	-72.83	-18.94
0.90	MCAR	-0.77	-24.28	-0.24
	MAR	-3.53	-67.99	-2.42
	MNAR	-9.84	-72.84	-9.32
			RMSE	
0.60	MCAR	0.0024	0.0123	0.0018
	MAR	0.0034	0.0269	0.0023
	MNAR	0.0146	0.0362	0.0143
0.75	MCAR	0.0018	0.0124	0.0015
	MAR	0.0028	0.0304	0.0021
	MNAR	0.0097	0.0362	0.0095
0.90	MCAR	0.0012	0.0124	0.0011
	MAR	0.0023	0.0339	0.0019
	MNAR	0.0051	0.0362	0.0048
			Coverage	
0.60	MCAR	0.65	0.06	0.73
	MAR	0.57	0.00	0.71
	MNAR	0.00	0.00	0.00
0.75	MCAR	0.68	0.06	0.74
	MAR	0.59	0.00	0.71
	MNAR	0.01	0.00	0.01
0.90	MCAR	0.74	0.06	0.75
	MAR	0.59	0.00	0.67
	MNAR	0.07	0.00	0.06

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table 5-6 displays the results when the size of the validation dataset was $0.20N$. The MNAR condition resulted in very large relative bias and the coverage became extremely small

even for Model 1 and Model 3. For Model 1 and Model 3, the coverage under the MAR condition was slightly smaller than the coverage under the MCAR condition. Model 2 performed badly when the size of the validation dataset was $0.20N$. And in this situation, Model 3 had a little higher coverage probability compared with Model 1. The RMSE for Model 1 was slightly larger than for Model 3, but the coverage of Model 1 was lower than that of Model 3 when the sensitivity was 0.60 or 0.75 under the MCAR condition. Under the MAR condition, the coverage of Model 1 was 16.28% less than that of Model 3. As for the effect of sensitivity of the observed disease status, the relative bias and RMSE of Model 1 decreased with increasing sensitivity. For Model 3, the relative bias and RMSE decreased with increasing sensitivity only for the MNAR condition. However, compared with the impact of the missingness mechanism, the impact of the sensitivity was not significant.

Table B-2 (Appendix B) displays the results of the Frequentist MI model (without bias correction) when the size of the validation dataset was $0.20N$. In this situation, Model 2 had the smallest relative bias and RMSE, and largest coverage compared with Model 1 and Model 3 of the Frequentist MI model under MCAR. The performance of Model 2 of the Frequentist MI approach was better than that of Model 2 of the Frequentist MI approach with bias correction. While the performance of Model 1 and Model 3 of the Frequentist MI approach were worse than the performance of Model 1 and Model 3 of the Frequentist MI approach with bias correction.

Table 5-7: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.35N$, Frequentist MI model with bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-2.88	-39.09	-0.12
	MAR	-4.27	-63.20	-1.45
	MNAR	-28.49	-78.48	-27.78
0.75	MCAR	-1.98	-39.06	-0.11
	MAR	-3.43	-69.06	-1.48
	MNAR	-18.49	-78.50	-18.04
0.90	MCAR	-0.82	-39.09	-0.12
	MAR	-2.50	-74.55	-1.45
	MNAR	-8.82	-78.45	-8.39
			RMSE	
0.60	MCAR	0.0021	0.0196	0.0016
	MAR	0.0027	0.0315	0.0019
	MNAR	0.0142	0.0390	0.0139
0.75	MCAR	0.0016	0.0196	0.0013
	MAR	0.0022	0.0344	0.0016
	MNAR	0.0093	0.0391	0.0090
0.90	MCAR	0.0010	0.0196	0.0009
	MAR	0.0017	0.0370	0.0013
	MNAR	0.0045	0.0390	0.0043
			Coverage	
0.60	MCAR	0.74	0.00	0.87
	MAR	0.66	0.00	0.85
	MNAR	0.00	0.00	0.00
0.75	MCAR	0.78	0.00	0.86
	MAR	0.67	0.00	0.84
	MNAR	0.00	0.00	0.00
0.90	MCAR	0.83	0.00	0.87
	MAR	0.70	0.00	0.81
	MNAR	0.02	0.00	0.02

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table 5-7 shows that the effect of the missingness mechanism was substantial when the size of the validation dataset was $0.35N$. Model 2 had very large relative bias and RMSE compared to

Model 1 and Model 3, similar to the results for the other sizes of the validation dataset. Model 3 had slightly larger coverage probability than Model 1. Increasing the sensitivity of observed disease status from 0.60 to 0.90 reduced the relative bias by 71.53% under the MCAR condition, by 51.45% under the MAR condition, and by 69.04% under the MNAR condition for Model 1, while Model 3 was stable with the increasing of the sensitivity of observed disease status under MCAR and MAR.

Table B-3 (Appendix B) reveals that the missingness mechanism was also substantial for the Frequentist MI model. For example, when the sensitivity was 0.90, the relative biases of Model 1 under MCAR, MAR and MNAR were -0.31%, -13.44% and 88.17%, respectively.

From Table 5-5 to 5-7, the Model 1 and Model 3 had smaller absolute values and RMSE with the size of the validation dataset increasing. And the coverage of Model 1 and Model 3 became larger when the size of the validation dataset increases only under MCAR and MAR. On the other hand, Model 2 had bigger absolute values of relative bias and RMSE and smaller coverage with the size of the validation dataset increasing. Tables in Appendix B show that the relative biases of all three models had no relationship with the size of the validation dataset. But the coverage of Model 2 significantly increased, and the coverage of Model 1 and Model 3 slightly decreased with the size of the validation dataset enlarger.

Table 5-8: Relative bias, RMSE and 95% confidence interval coverage for different conditions of measurement error in covariates when the size of the validation dataset is $0.05N$, Frequentist MI model with bias correction

Additive Variance	Misclassification	Model 1								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	1.53	-5.03	-20.22	0.0031	0.0045	0.0107	0.58	0.55	0.21
	SN=SP=0.70	1.40	-5.48	-20.36	0.0031	0.0045	0.0108	0.58	0.54	0.20
2	SN=SP=0.90	-0.96	-8.03	-21.71	0.0030	0.0055	0.0115	0.56	0.49	0.18
	SN=SP=0.70	-0.80	-7.93	-21.76	0.0030	0.0054	0.0115	0.56	0.49	0.18
		Model 2								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-1.99	-45.78	-59.91	0.0044	0.0230	0.0299	0.53	0.05	0.01
	SN=SP=0.70	-2.01	-45.89	-60.12	0.0044	0.0231	0.0301	0.52	0.05	0.01
2	SN=SP=0.90	-6.17	-49.17	-62.19	0.0054	0.0247	0.0311	0.49	0.04	0.01
	SN=SP=0.70	-6.16	-49.03	-62.07	0.0053	0.0246	0.0310	0.50	0.04	0.01
		Model 3								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-0.55	-6.42	-22.93	0.0027	0.0047	0.0119	0.56	0.50	0.13
	SN=SP=0.70	-0.42	-6.64	-22.91	0.0027	0.0048	0.0119	0.55	0.51	0.12
2	SN=SP=0.90	-0.41	-6.68	-22.67	0.0026	0.0048	0.0118	0.55	0.50	0.13
	SN=SP=0.70	-0.35	-6.48	-22.67	0.0027	0.0047	0.0118	0.55	0.50	0.13

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. SN = sensitivity; SP = specificity. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Next, the effects of the magnitude of the error in the covariates on the performance of three different models were described. As Table 5-8 indicates, Model 1 had slightly larger coverage probability (8.12% on average) when the additive variance Var(ε_1)=1 than when the additive variance Var(ε_1)=2, regardless of the missingness mechanism. Model 1 had slightly less absolute value of relative bias and RMSE when the additive variance Var(ε_1)=1 than when the additive variance Var(ε_1)=2. For Model 2, when the additive variance Var(ε_1)=1, the relative bias and RMSE were slightly smaller and the coverage probability was slightly greater than when the additive variance Var(ε_1)=2 under the MCAR and MAR conditions. Both Model 1 and Model 2 were not sensitive to the different values of the sensitivity and specificity of the binary covariate. And for the Model 3, the results were similar across different magnitudes of the measurement error in the covariates.

Table 5-9: Relative bias, RMSE and 95% confidence interval coverage for different conditions of measurement error in covariates when the size of the validation dataset is $0.20N$, Frequentist MI model with bias correction

Additive Variance	Misclassification	Model 1								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-1.35	-2.86	-19.02	0.0016	0.0020	0.0096	0.72	0.74	0.03
	SN=SP=0.70	-1.29	-2.85	-19.12	0.0016	0.0020	0.0096	0.73	0.74	0.03
2	SN=SP=0.90	-2.63	-3.92	-19.68	0.0020	0.0024	0.0099	0.66	0.62	0.02
	SN=SP=0.70	-2.74	-3.96	-19.79	0.0020	0.0024	0.0100	0.65	0.62	0.02
		Model 2								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-22.89	-68.08	-72.06	0.0117	0.0339	0.0358	0.07	0.00	0.00
	SN=SP=0.70	-22.71	-68.13	-72.18	0.0116	0.0339	0.0359	0.08	0.00	0.00
2	SN=SP=0.90	-25.70	-69.75	-73.42	0.0131	0.0347	0.0365	0.05	0.00	0.00
	SN=SP=0.70	-25.69	-69.80	-73.48	0.0130	0.0347	0.0366	0.04	0.00	0.00
		Model 3								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-0.29	-1.49	-18.99	0.0015	0.0016	0.0095	0.74	0.83	0.02
	SN=SP=0.70	-0.19	-1.44	-19.00	0.0015	0.0016	0.0096	0.75	0.83	0.02
2	SN=SP=0.90	-0.13	-1.46	-18.92	0.0015	0.0016	0.0095	0.74	0.83	0.02
	SN=SP=0.70	-0.30	-1.46	-18.94	0.0015	0.0016	0.0095	0.74	0.83	0.02

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. Model 3 is the predictive model with the observed response only. SN = sensitivity; SP = specificity. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

The results shown in Table 5-9 were based on the validation dataset whose size was $0.20N$. Under the MCAR, MAR and MNAR conditions and when the additive variance Var(ε_1)=1, the coverage probability of Model 1 was 10.69%, 19.35% and 50.00% greater than the coverage probability when the additive variance Var(ε_1)=2, respectively. For example, under the MAR condition when the additive variance Var(ε_1)=1 and sensitivity and specificity were both equal to 0.90, the relative bias, RMSE and coverage probability were -2.86%, 0.0020 and 0.74, respectively. Under the MAR condition, when the additive variance Var(ε_1)=1 and sensitivity and specificity were both equal to 0.70, the relative bias, RMSE and coverage probability were -2.85%, 0.0020 and 0.74, respectively. However, under the MAR condition when the additive variance Var(ε_1)=2 and sensitivity and specificity were equal to 0.90, the relative bias, RMSE and coverage probability were -3.92%, 0.0024 and 0.62, respectively. For Model 2, when the additive variance Var(ε_1)=2, the average of the absolute relative bias over three the missingness mechanisms was 3.62% greater than that when the additive variance Var(ε_1)=1. However, the coverage probability of Model 2 was very small and does not change as the magnitude of measurement error in the covariates change. Model 3 still had stable values of relative bias, RMSE, and coverage probability with the additive variance increases and sensitivity and specificity decrease. For example, under the MAR condition the coverage was 0.83 regardless of the amount of error in the continuous covariate and the sensitivity and specificity of the binary covariate.

Table 5-10: Relative bias, RMSE and 95% confidence interval coverage for different conditions of measurement error in covariates when the size of the validation dataset is $0.35N$, Frequentist MI model with bias correction

Additive Variance	Misclassification	Model 1								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-1.37	-2.86	-18.36	0.0014	0.0020	0.0092	0.82	0.74	0.01
	SN=SP=0.70	-1.38	-2.85	-18.34	0.0014	0.0020	0.0092	0.82	0.74	0.01
2	SN=SP=0.90	-2.41	-3.92	-18.84	0.0017	0.0024	0.0094	0.74	0.62	0.01
	SN=SP=0.70	-2.41	-3.96	-18.86	0.0017	0.0024	0.0094	0.74	0.62	0.01
		Model 2								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-37.95	-68.08	-77.96	0.0190	0.0339	0.0388	0.00	0.00	0.00
	SN=SP=0.70	-37.87	-68.13	-78.00	0.0190	0.0339	0.0388	0.00	0.00	0.00
2	SN=SP=0.90	-40.27	-69.75	-78.93	0.0202	0.0347	0.0393	0.00	0.00	0.00
	SN=SP=0.70	-40.22	-69.80	-79.02	0.0201	0.0347	0.0393	0.00	0.00	0.00
		Model 3								
		Relative Bias (%)			RMSE			Coverage		
		MCAR	MAR	MNAR	MCAR	MAR	MNAR	MCAR	MAR	MNAR
1	SN=SP=0.90	-0.14	-1.49	-18.08	0.0013	0.0016	0.0091	0.87	0.83	0.01
	SN=SP=0.70	-0.12	-1.44	-18.07	0.0013	0.0016	0.0091	0.87	0.83	0.01
2	SN=SP=0.90	-0.07	-1.46	-18.09	0.0013	0.0016	0.0091	0.87	0.83	0.01
	SN=SP=0.70	-0.14	-1.46	-18.06	0.0013	0.0016	0.0090	0.87	0.83	0.01

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. SN = sensitivity; SP = specificity. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table 5-10 reveals that when the size of the validation dataset increased to $0.35N$ the magnitude of the measurement error in the covariates remained influential for Model 1 but has less impact on the results for Model 2. Specifically, under the MCAR condition, when the additive variance Var(ε_1)=1, the coverage was 10.81% larger than when the additive variance Var(ε_1)=2. And under the MAR condition when the additive variance Var(ε_1)=1, the coverage was 19.35% greater than when the additive variance Var(ε_1)=2. For Model 2, there was little difference in the relative bias and RMSE between the two measurement error conditions for the continuous covariate, but the coverage for Model 2 were all close to zero. As for Model 3, it still remained stable across different magnitude of the measurement error in covariates.

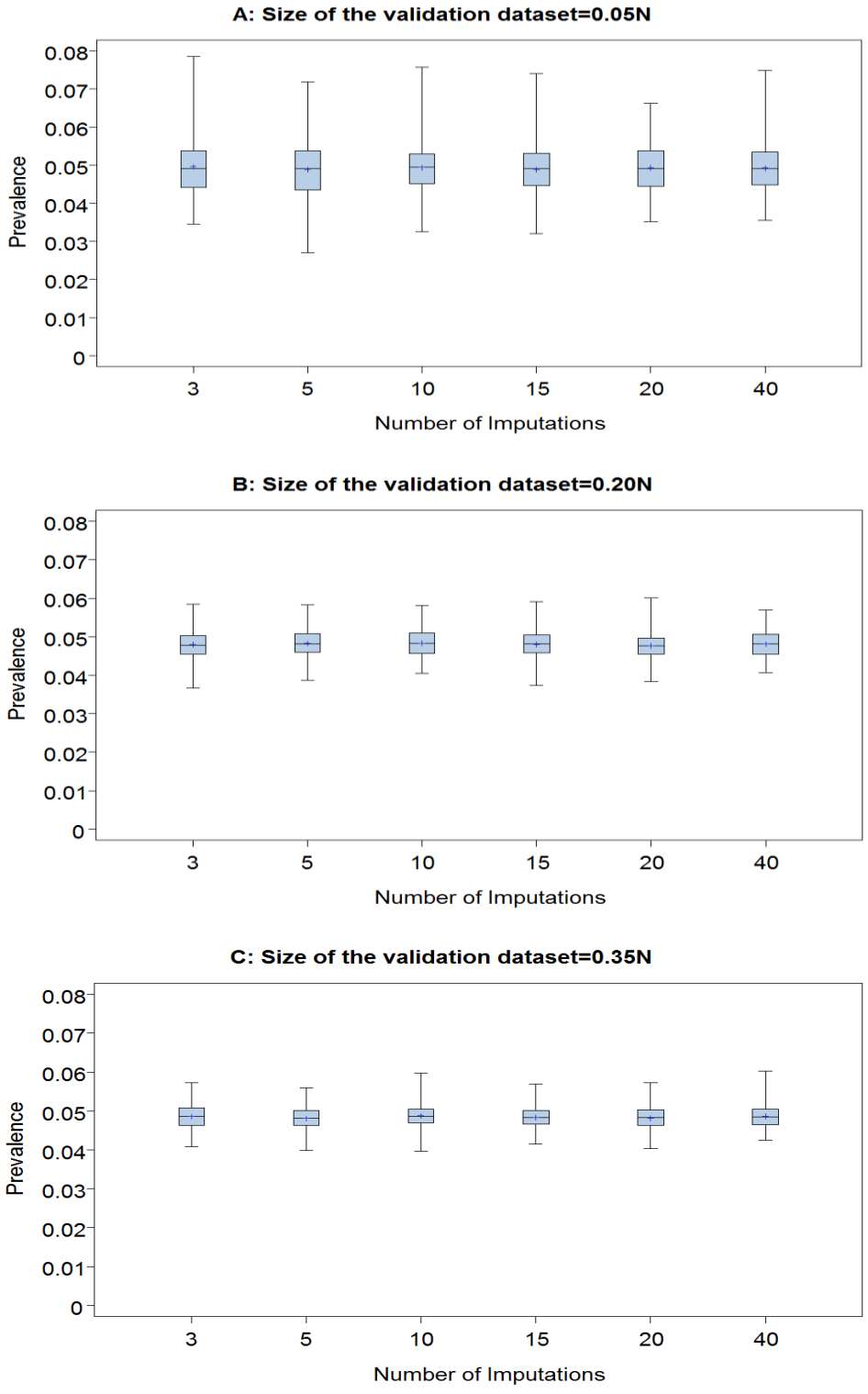

Figure 5-7 Prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MCAR

The effect of the number of imputations is examined next. Figure 5-7 reveals that the average prevalence estimate remains close to 0.05 regardless the size of the validation dataset. In panel A, the prevalence estimates range from 0.03 to 0.08 and does not change with the number of imputations. Specifically, the minimum and maximum of the prevalence estimates with three imputations are 0.030 and 0.076 (mean is 0.049) and the minimum and maximum of the prevalence estimates with forty imputations are 0.034 and 0.066 (mean is 0.050). Panel C displays the range of prevalence estimates when the size of the validation dataset is $0.35N$. The range is similar for different number of imputations. For example, the minimum and maximum of the prevalence estimates with three imputations are 0.042 and 0.058 (mean is 0.048) and the minimum and maximum of the prevalence estimates with forty imputations are 0.040 and 0.056 (mean is 0.048). Comparing panel A and panel C, we can notice that the range of the prevalence estimates based on $0.35N$ validation dataset is eloquently smaller than that based on $0.05N$ validation dataset through all different numbers of imputations. But the mean of the prevalence estimates over 500 replications for specific number of imputations remains close to the true prevalence.

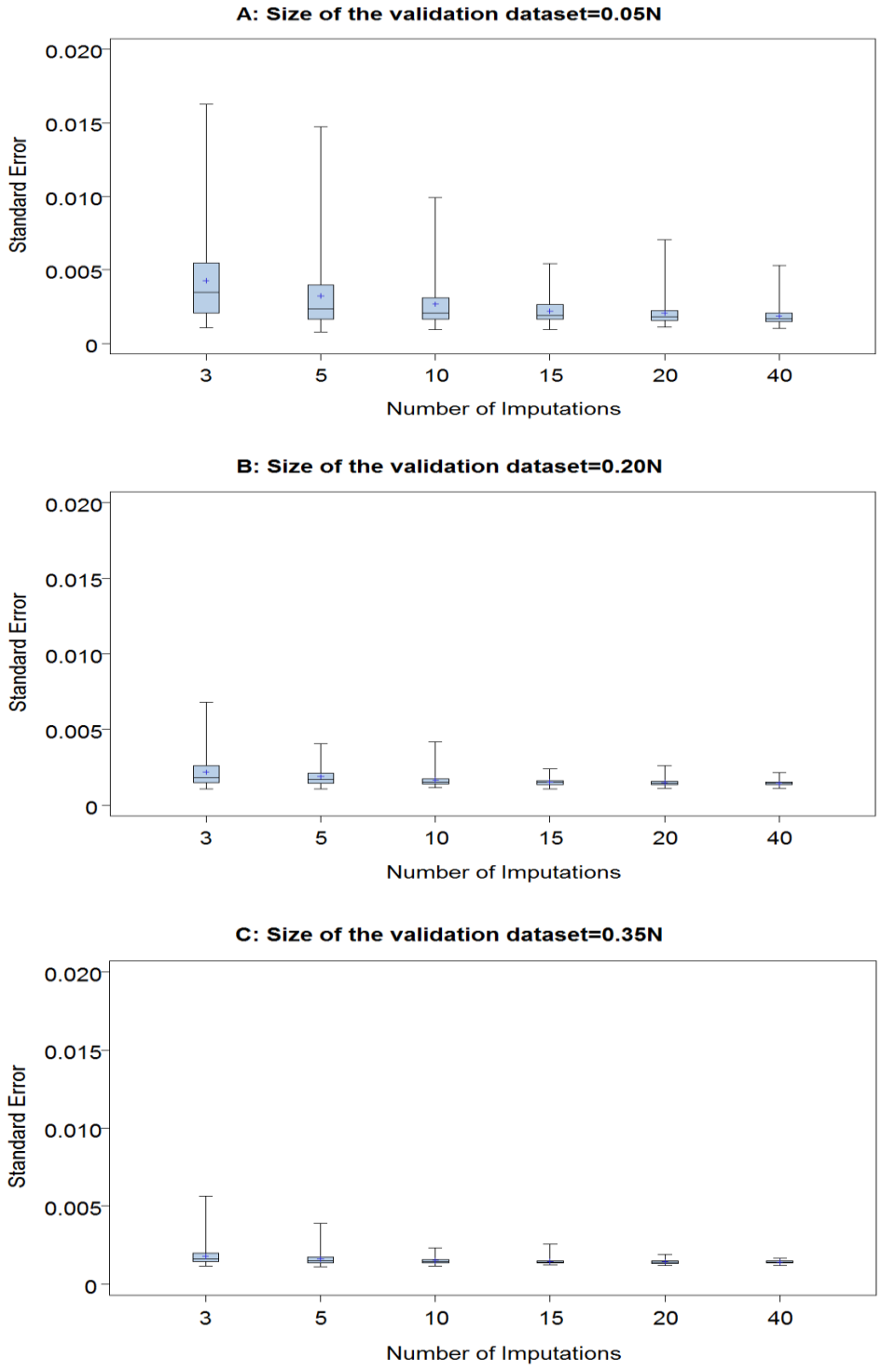

Figure 5-8 Standard errors of prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MCAR

As Figure 5-8 reveals, the range of the standard error of the prevalence estimate decrease from almost 0.020 to just over 0.005 when the number of imputations increases from 3 to 15. However, when the number of imputations is 20, the range increases, and is similar in size to when the number of imputations is three. When the number of imputations is 40, the range of the standard error of prevalence estimate decreases to approximately 0.005. The mean of the standard error of prevalence estimate decreased by 45.87% as the number of imputations increases from three to 10. Panel C of Figure 5-8 shows the box plot of the standard error of prevalence estimate when the size of the validation dataset is $0.35N$. In this situation, the mean of the standard error of prevalence estimate slightly decrease from 0.002 to 0.001 and the range also decreases with the number of imputations increases except than when the number of imputations is 10.

For the MCAR condition, the range and mean of the prevalence estimates was constant across the different dataset size conditions. As for the range and mean of the standard error of the prevalence estimates they decreased significantly as the number of imputations increased when the size of the validation dataset was $0.05N$ and decrease slightly when the size of the validation data set was $0.35N$.

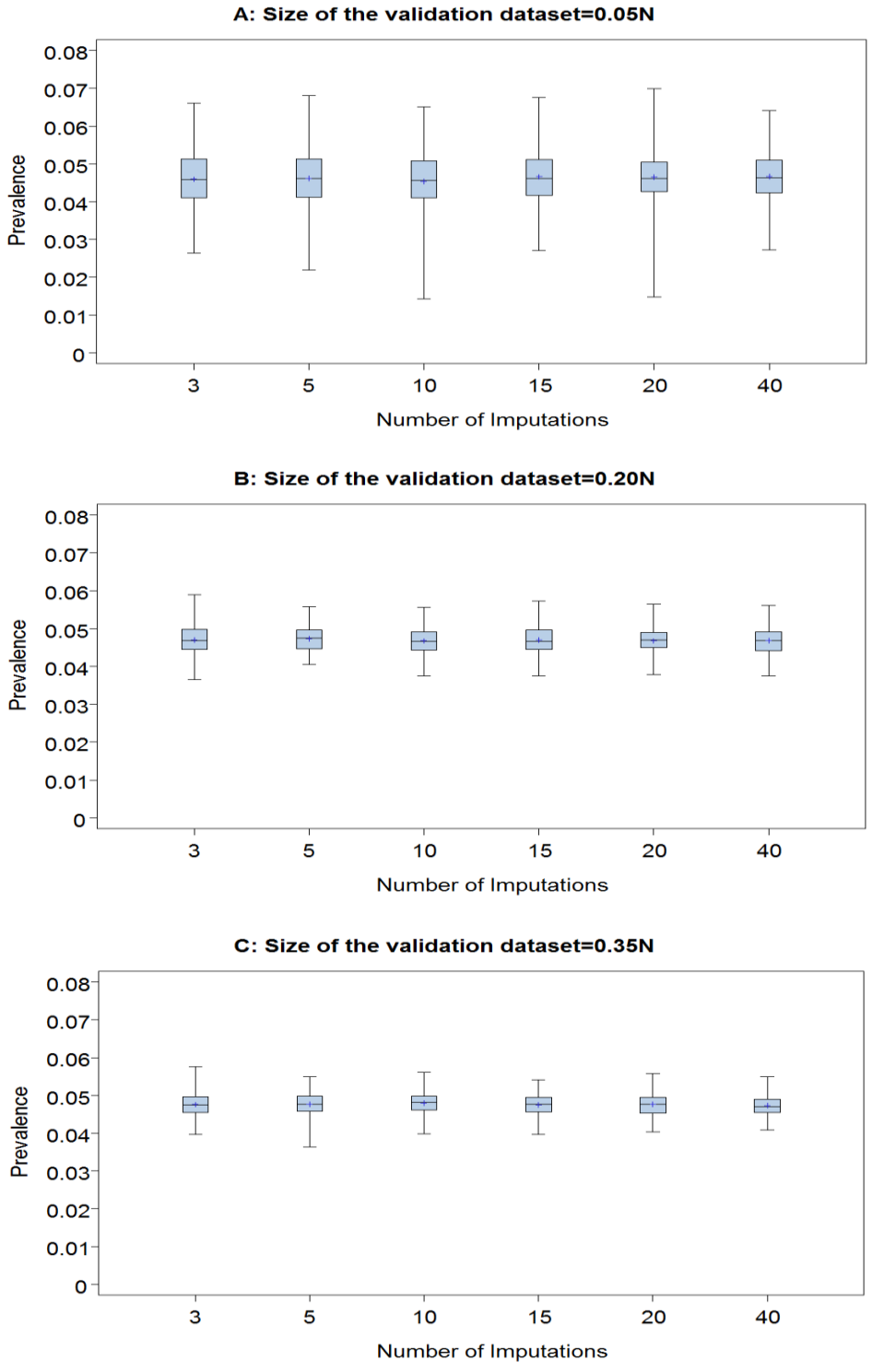

Figure 5-9 Prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MAR

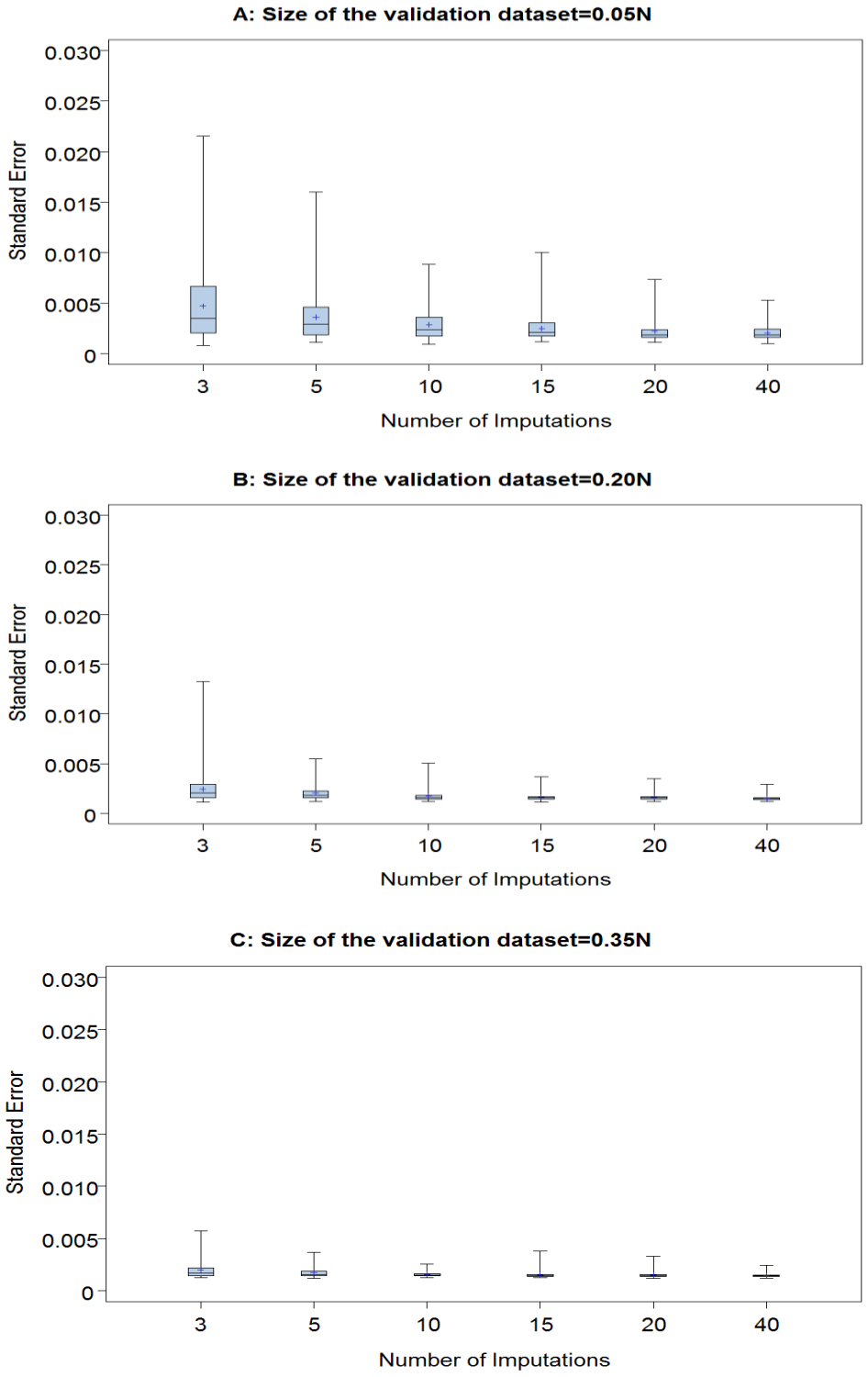

Figure 5-10 Standard errors of prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MAR

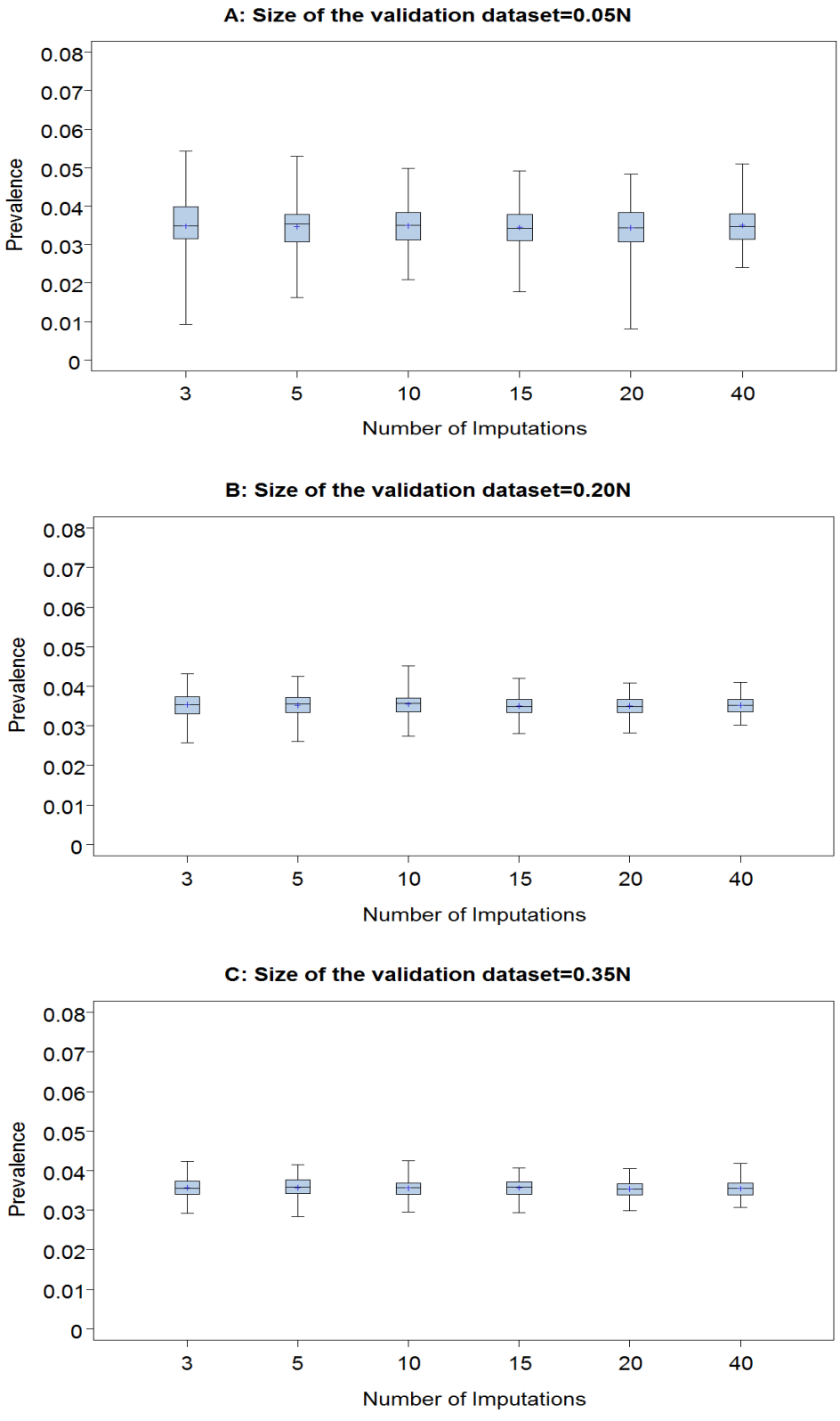

Figure 5-11 Prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MNAR

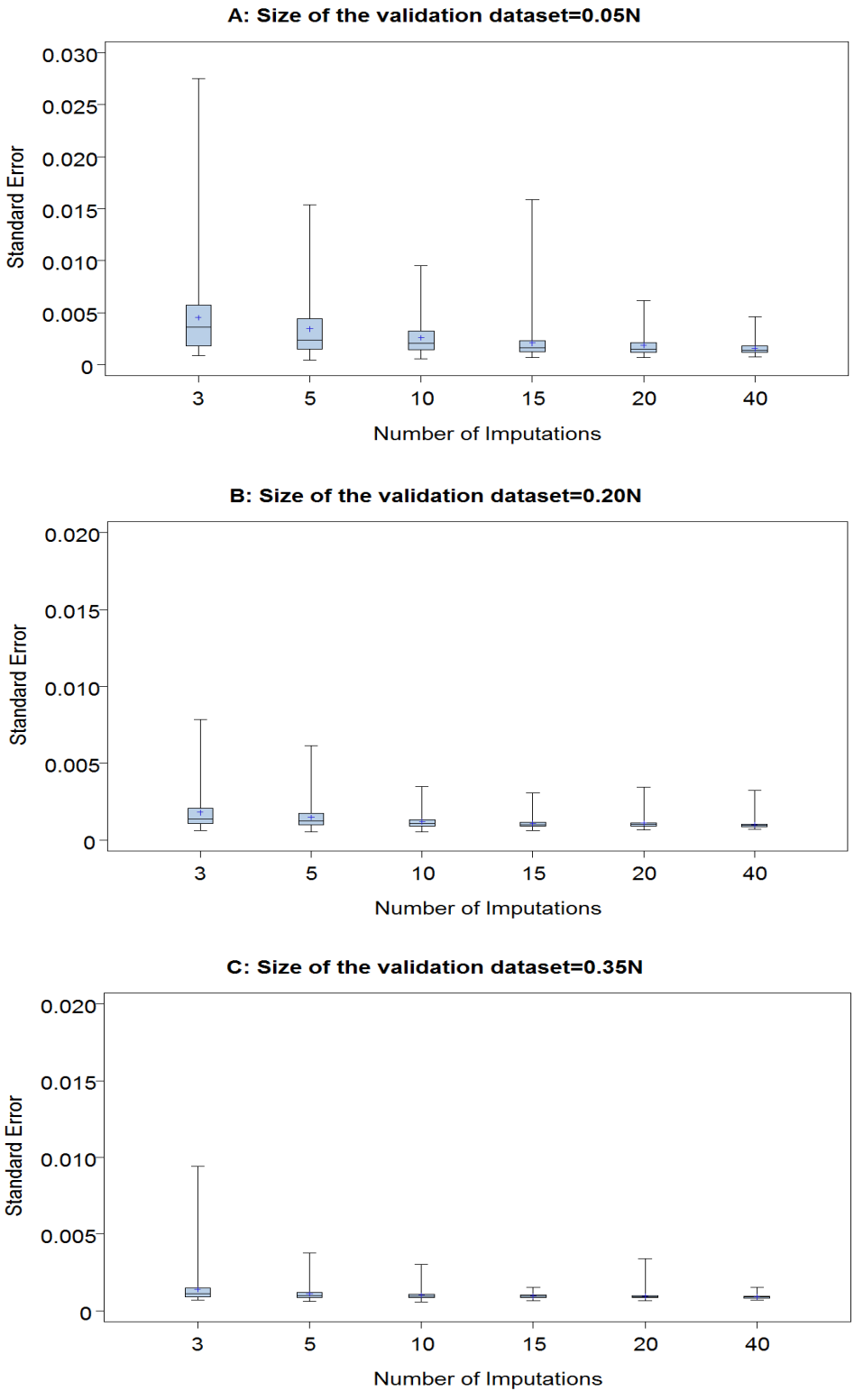

Figure 5-12 Standard errors of prevalence estimates of Model 3 by number of imputations when the sensitivity is 0.60, Frequentist MI model with bias correction, and the missingness mechanism is MNAR

Other than the figures shown above, the trend of the prevalence estimates and standard errors of prevalence estimates with increasing size of the validation dataset and the number of imputations for other values of sensitivity and models were same as when the sensitivity was 0.60 and for Model 3.

CHAPTER 6. CONCLUSIONS AND DISCUSSION

The purpose of the research was to investigate different MI methods that can estimate the disease prevalence accurately and examine the factors that influence the performance of MI methods. The simulation study fulfilled the objectives that MI methods using Frequentist or Bayesian logistic regression model as predictive model could improve the prevalence estimate; the characteristics of model-based case-detection algorithms including sensitivity of observed disease status and type and magnitude of measurement error in the predictors of disease status slightly influenced the performance of the MI methods; and the effect of the size of the validation dataset was significant to accurately estimate the prevalence while increasing the number of imputations on the performance of MI methods only decreased the standard error of the prevalence estimate.

6.1 Conclusions

The Frequentist MI model with bias correction had superior performance of estimating disease prevalence to Frequentist MI model without bias correction and Bayesian MI model regardless the size of the validation dataset, which was demonstrated via the simulation study in scenario 1. In scenario 1, the Frequentist MI model with bias correction performed well under MCAR and MAR with small relative bias, RMSE and reasonable coverage probability of the 95% confidence interval. When the sensitivity was 0.75 or 0.90, the Bayesian MI model had smaller relative bias and RMSE than the Frequentist MI model without bias correction, but under MCAR the Bayesian MI model had substantially smaller coverage compared with the Frequentist MI model without bias correction. For the Bayesian MI model, the relative bias under MNAR was greater than the relative bias under MCAR and MAR, and the relative bias decreased with the size of the vlidation dataset increasing. For the Frequentist MI model without bias correction, the relative bias and RMSE increased with the size of the validation increasing.

In both scenario 1 and scenario 2, the missingness mechanism had a significant impact on the accuracy of the prevalence estimates. Especially, when the missingness mechanism was MNAR that is the data violated the MAR assumption of the MI methods, even the Frequentist MI model with bias correction would have very small coverage.

Based on the results of scenario 1 and scenario 2, it is demonstrated that the predictive model with observed disease status as covariate can achieve the goal to estimate the disease prevalence correctly. In scenario 1, the Frequentist MI model with bias correction having observed disease status as covariate of the predictive model had the smallest relative bias regardless of the missingness mechanism, the size of the validation dataset and the sensitivity of the observed disease status. In scenario 2, the Frequentist MI model with bias correction having observed disease status as covariate of the predictive model also had the smallest relative bias compared with the other two predictive models. Table 5-5, Table 5-6 and Table 5-7 suggest that the Model 1 with the observed disease status and imputed disease predictors as covariates and Model 3 with observed disease status only can be applied to estimate the prevalence. And the model (Model 2) only with imputed disease predictors as covariates has low coverage probability indicating that the method to impute the disease predictors first and then use the imputed disease predictors to impute the true disease status did not perform well, especially when the missingness mechanism is MAR. From Table 5-8, Table 5-9 and Table 5-10, one can quickly make the observation that larger additive variance in the continuous disease predictor can result in bigger relative bias, RMSE and smaller coverage than when variance is smaller in the disease predictor for Model 1. Generally, the Model 3 did not change with the change of the magnitude of the measurement error in disease predictors. The reason is that Model 3 only has the observed disease status as the covariate of predictive model. However, if the Frequentist MI model without bias correction is used, Model 2 performed better than Model 1 and Model 3.

The effect of the size of the validation dataset was significant both in scenario 1 and scenario 2. Increasing the size of the validation dataset reduced the relative bias and RMSE for both Frequentist MI model with bias correction and Bayesian MI model as well as for different predictive models under different missingness mechanisms. This is reasonable because the larger the size of the validation dataset is the more information are provided to build the model which can correctly reflect the association between the observed disease status and true disease status. It is recommended to increase the size of the validation dataset when implementing the MI methods to correct for measurement errors.

The consequence of the sensitivity of the observed disease status depends on the MI method and the missingness mechanism. Generally, if the missingness mechanism is MNAR, the relative bias and RMSE would decrease with the sensitivity of the observed disease status increasing. In

scenario 1, it is also shown that the Bayesian MI model had smaller relative bias and RMSE when the sensitivity of the observed disease status is increased.

The results for scenario 1 and scenario 2 also demonstrate that a large number of imputations can improve the precision of the prevalence estimate regardless the missingness mechanism. The range and mean of standard errors for different validation dataset sizes and different missingness mechanisms (including MCAR, MAR and MNAR) all decreased in varying degrees with the number of imputations increasing. Overall, we can see that the effect of the number of imputations was important when the size of the validation dataset was $0.05N$, while it became less important when the size of the validation dataset was as large as $0.35N$, regardless of the missingness mechanism. Increase the validation dataset can improve both the accuracy and precision of prevalence estimate, and increase the number of imputations can only improve the precision.

6.2 Discussion

The predictive models in this research were constructed using a logistic regression model with selected variables. In medical research, the logistic regression model is widely used to construct predictive models[90,91]. By using logistic regression, the observed disease status is directly modeled to the true disease status. In this way, the MI can be easily implemented to impute the individual's true disease status so that the prevalence of the entire population is estimated. On the other hand, the predictive model can also be defined by the conditional distribution of true disease status on all the information provided by observed disease status and disease predictors. The alternative approach is sequential regression multiple imputation, which specifies a series of separate conditional distributions for each incomplete variable without the need to fit a multivariate model[92]. In this way, the relationship between the observed disease status and true disease status is studied based on the separate models, which preserves the structure in the data as well as the uncertainty about this structure and includes any knowledge about the process that generated the missing data. Usually, these foundational models are specified by expertize in the field, which may also be misspecified. We also studied the sequential regression multiple imputation that can be used as the predictive models in Frequentist approach and as the joint distribution in fully Bayesian approach (Appendix C). The sequential

regression models include the model of sensitivity and specificity (i.e., in scenario 1 is the true disease status conditional on the observed disease status; in scenario 2 is the true disease status conditional on the observed disease status and disease predictors), the disease model, the measurement error model for continuous disease predictor, and the misclassification model for binary disease predictor.

In scenario 1, the comparison of the Bayesian MI model and the Frequentist MI model without bias correction substantiated the theory that the Bayesian approach with non-informative prior distribution should have approximate results as the maximum likelihood estimation of Frequentist approach. However, in reality the size of the validation dataset is quite small and the prevalence is very low. That means we encountered the problem of small sample size and sparse data (i.e., data with few or no subjects at crucial combinations of variable values). For the Frequentist MI model without bias correction, the relative bias and RMSE were the largest when the sensitivity was 0.90 under the MNAR condition. The reason is the very small sample size and the unbalanced values in the contingency table with sensitivity 0.9 and specificity 1.0. We implemented the bias correction algorithm with maximum likelihood estimation in Frequentist approach and it produced small relative bias, good precision and confidence coverage. The Bayesian approach may be improved by selecting an appropriate prior distribution based on expert recommendation. A number of studies have developed different informative prior distributions that can also improve the inference of Bayesian approach[93,94].

In scenario 2, we investigated predictive models with different selected variables as covariates of the predictive model based on Frequentist MI model with bias correction. It is generally recommended to include a rich set of predictors that are relevant to the imputed variable when imputing missing values[95]. However, the results indicate that the observed disease status as covariate of the predictive model to impute the true disease status is as good as the predictive model with observed disease status and imputed disease predictors. If the primary object of this research is to correctly estimate the association between the disease predictors and the disease status, using the predictive model with observed disease status and imputed disease predictors to impute true disease status should be helpful to correctly estimate the coefficients of the disease model.

The assumption of the MI methods in this research was that non-differential measurement error exists in the covariates, which means the measurement error models are the same for individuals with disease as those without disease. In AHDs, the data are usually collected prospectively, thus the non-differential measurement error assumption of the covariates is assumed to be reasonable. In addition, the non-differential measurement error assumption is appropriate, except for case-control studies and when $\boldsymbol{W} = (W_1, W_2)$ is not the observed measure of $\boldsymbol{X} = (X_1, X_2)$ but a surrogate variable as a proxy of $\boldsymbol{X}$. That means for case-control studies the measure of the true values tend to be biased. Thus the methods here may not be applied to problems in which measurement error models are different for individuals with and without disease.

If the missingness mechanism is MNAR, MI methods could not substantially improve the accuracy of population disease prevalence estimates over the naïve estimation method. So it is important to identify the missingness mechanism based on how the validation study was conducted. In one study, the diagnoses contained in the medical charts of the cohort of Quebec seniors enrolled in the medical office of the 21st century (MOXXI) study were used to validate administrative health data[5]. This validation dataset should result in the MCAR mechanism for missing records of people not enrolled in the MOXXI study because the study adopted a randomized trial design in a study population of 110 primary care physicians in Montreal, Quebec. In a different study, Bone Mineral Density (BMD) testing results from Manitoba were used to assess the validity of osteoporosis cases ascertained from administrative databases.[11] It is likely that the MAR mechanism characterizes the missing test results. Bone Mineral Density (BMD) testing in Manitoba requires physician referral; individuals who are referred are more likely to be female and age 65 years or older. Thus, whether a person is in the testing program is conditional on observed covariates. Another study to validate the accuracy of identification of patients with immune thrombocytopenic purpura (ITP) through administrative records[38] is likely to result in MNAR missing diagnoses from the Electronic Patient Record (the medical record) of people who do not have the records. The reason is that the records were retrieved from inpatients and outpatients with ITP as a primary or secondary diagnosis which was missing if people do not have the medical record of ITP.

6.3 Summary and Recommendations

Our finding suggest that using MI methods to improve the accuracy of case ascertainment in AHDs can result in a valid dataset for public usage including surveillance, utilization review of services, policy evaluation, and risk adjustment. In this research, the MI methods accounted for the measurement error both in disease status and disease predictors, which appears frequently in practice. More reliable estimates of variability are provided by accommodation of the uncertainty caused by missing data. So it is recommended to apply MI methods to correct the potentially inaccurate disease diagnoses in AHDs before using the AHDs to do further research or decision-making.

We examined MI methods using both a Bayesian and Frequentist logistic model, which, to the best of our knowledge, has not previously been studied. Considering the characteristics of the AHD, we developed a valid MI method (in Frequentist approach with Bias Correction algorithm) that can estimate the disease prevalence even though the validation dataset is small and the specificity is close to one.

From the results, the Frequentist MI with bias correction is not sensitive to the sensitivity of the observed disease status under MCAR and MAR missingness mechanisms. Thus the MI methods can be applied to address the misclassification of the observed response when it is assumed to be under-reported.

The investigation of the potential factors that may influence the performance of the estimation of the disease prevalence provides a guideline for future research to deal with the missing data and measurement error problem. The most appropriate method can be implemented depending on the missingness mechanism and the size of the validation dataset. If the sampling mechanism of the validation dataset can be expected to result in MCAR or MAR missing data in the main dataset, the size of the validation dataset is less than 35% of the population, and there is no specified prior distribution for the Bayesian method, then the Frequentist MI model with bias correction provides more accurate estimates of prevalence and is recommended based on our simulation study. When selecting covariates for the predictive model, regardless of whether the misclassification of the observed disease status is dependent on or independent of disease predictors that contain measurement error, the Frequentist MI model with bias correction is

recommended. And it is recommended that when applying the MI method, the number of imputations toned not be greater than 20. Not only will more imputations result in a longer computational time, but may also decrease the coverage probability (i.e., precision) of the confidence intervals.

6.4 Future Research

There are a number of opportunities for future research. The study is based on the main/validation dataset design in which the validation dataset contains disease status and disease predictors measured without error. However, the validation dataset may also contain mismeasured values. Further research might explore this situation. Methods for measurement error correction with the assumption that the 'true' values gathered from validation dataset are also error-prone should be investigated.

We compared the MI methods based on Frequentist logistic model and Bayesian logistic model in scenario 1. We would like to explore the MI method using Bayesian logistic model in scenario 2 in next step. The MI method using Bayesian logistic model is using MCMC to draw the samples of the parameters of the predictive logistic model, then the multiple values of true disease status are imputed using the multiple values of the parameters. The fully Bayesian MI method that draw the multiple samples of the disease status from the posterior distribution based on the joint distribution also deserves investigation so that it can be applied in AHDs.

In this research, we propose the method to correctly ascertain the disease cases by taking use of the observed disease status and relevant disease markers. For future research, the association between the true disease status and the disease predictors might be examined. We will examine the performance of MI methods to estimate the coefficient of the covariates in the disease model.

The attention of the study has been restricted to the case where the covariates of the disease model are independent. The disease markers maybe correlated with each other. Thus extensions to the correlated covariates could also be developed.

In practice, the MAR assumption is very difficult to test. And the MNAR is possible, so the development of the method that can be implemented to deal with the missing data even the missingness mechanism is MNAR would be valuable and innovative.

REFERENCES

1. Hux JE, Ivis F, Flintoft V, Bica A. Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care.* 2002; 25(3): 512-516.

2. Maskarinec G. Diabetes in Hawaii: estimating prevalence from insurance claims data. *American Journal of Public Health.*1997; 87(10): 1717-1720.

3. Rector TS, Wickstrom SL, Shah M, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. *Health Services Research.* 2004; 39(6 Pt 1): 1839-1857.

4. Lix L, Yogendran M, Burchill C, et al. *Defining and validating chronic diseases: An administrative data approach*. Winnipeg: Manitoba Centre for Health Policy; 2006.

5. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. *Journal of Clinical Epidemiology.* 2004; 57(2): 131-141.

6. Tu K, Campbell NR, Chen ZL, Cauch-Dudek KJ, McAlister FA. Accuracy of administrative databases in identifying patients with hypertension. *Open Medicine.* 2007; 1(1): e18-26.

7. Quan H, Khan N, Hemmelgarn BR, et al. Validation of a case definition to define hypertension using administrative data. *Hypertension.* 2009; 54(6): 1423-1428.

8. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare's hospital claims data: progress has been made, but problems remain. *American Journal of Public Health.*1992; 82(2): 243-248.

9. Friis RH, Sellers TA. *Epidemiology for Public Health Practice.* Sudbury: Jones and Bartlett; 2009.

10. Vestergaard P, Rejnmark L, Mosekilde L. Osteoporosis is markedly underdiagnosed: a nationwide study from Denmark. *Osteoporos International.* 2005; 16(2): 134-141.

11. Lix LM, Yogendran MS, Leslie WD, et al. Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. *Journal of Clinical Epidemiology.* 2008; 61(12): 1250-1260.

12. Ladouceur M, Rahme E, Pineau CA, Joseph L. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics.* 2007; 63(1): 272-279.

13. Molodecky NA, Myers RP, Barkema HW, Quan H, Kaplan GG. Validity of administrative data for the diagnosis of primary sclerosing cholangitis: a population-based study. *Liver International.* 2011; 31(5): 712-720.

14. Carroll RJ. Measurement error in epidemiologic studies. *Encyclopedia of Biostatistics*. New York: Wiley; 2005: 2491-2519.

15. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology.* 2006; 35(4): 1074-1081.

16. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition.* Boca Raton: Taylor & Francis; 2006.

17. Schenker N, Raghunathan TE, Bondarenko I. Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics Medicine.* 2010; 29(5): 533-545.

18. Barnard J, Rubin DB. Small-Sample degrees of freedom with multiple imputation. *Biometrika.* 1999; 86(4): 948-955.

19. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *British Medical Journal.* 2009; 338: b2393.

20. Holan SH, Toth D, Ferreira MAR, Karr AF. Bayesian multiscale multiple imputation with implications for data confidentiality. *Journal of the American Statistical Association.* 2010; 105(490): 564-577.

21. Raghunathan T. Combining information from multiple surveys for assessing health disparities. *Allgemeines Statistisches Archive.* 2006; 90(4): 515-526.

22. Padilla MA, Divers J, Vaughan LK, Allison DB, Tiwari HK. Multiple imputation to correct for measurement error in admixture estimates in genetic structured association testing. *Humn Heredity.* 2009; 68(1): 65-72.

23. Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments.* Boca Raton: Chapman & Hall/CRC; 2003.

24. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley; 1987.

25. Kipnis V, Subar AF, Midthune D, et al. Structure of dietary measurement error: results of the OPEN biomarker study. *American Journal of Epidemiology.* 2003; 158(1): 14-21; discussion 22-16.

26. Neuhouser ML, Tinker L, Shaw PA, et al. Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. *American Journal Epidemiology.* 2008; 167(10): 1247-1259.

27. Kopecky KJ, Davis S, Hamilton TE, Saporito MS, Onstad LE. Estimation of thyroid radiation doses for the hanford thyroid disease study: results and implications for statistical power of the epidemiological analyses. *Health Physics.* 2004; 87(1): 15-32.

28. Vasan RS, Massaro JM, Wilson PWF, et al. Antecedent blood pressure and risk of cardiovascular disease the Framingham heart study. *Circulation.* 2002; 105(1): 48-53.

29. Gordon T, Kannel WB. Premature mortality from coronary heart disease the Framingham study. *Journal of the American Medical Association.* 1971; 215(10): 1617-1625.

30. MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet.* 1990; 335(8692): 765-774.

31. Malenka DJ, McLerran D, Roos N, Fisher ES, Wennberg JE. Using administrative data to describe casemix: a comparison with the medical record. *Journal of Clinical Epidemiology.* 1994; 47(9): 1027-1032.

32. Muhajarine N, Mustard C, Roos LL, Young TK, Gelskey DE. Comparison of survey and physician claims data for detecting hypertension. *Journal of Clinical Epidemiology.* 1997; 50(6): 711-718.

33. Petersen LA, Wright S, Normand SL, Daley J. Positive predictive value of the diagnosis of acute myocardial infarction in an administrative database. *International Journal of General Medicine.* 1999; 14(9): 555-558.

34. Quan H, Parsons GA, Ghali WA. Validity of procedure codes in International Classification of Diseases, 9th revision, clinical modification administrative data. *Medical Care.* 2004; 42(8): 801-809.

35. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. *Journal of Clinical Epidemiology.* 2011; 64(10): 1054-1059.

36. Newton KM, Wagner EH, Ramsey SD, et al. The use of automated data to identify complications and comorbidities of diabetes: a validation study. *Journal of Clinical Epidemiology.* 1999; 52(3): 199-207.

37. Quan H, Parsons GA, Ghali WA. Validity of information on comorbidity derived rom ICD-9-CCM administrative data. *Medical Care.* 2002; 40(8): 675-685.

38. Segal JB, Powe NR. Accuracy of identification of patients with immune thrombocytopenic purpura through administrative records: a data validation study. *American Journal of Hematology.* 2004; 75(1): 12-17.

39. Klabunde CN, Potosky AL, Legler JM, Warren JL. Development of a comorbidity index using physician claims data. *Journal of Clinical Epidemiology.* 2000; 53(12): 1258-1267.

40. Humphries KH, Rankin JM, Carere RG, Buller CE, Kiely FM, Spinelli JJ. Co-morbidity data in outcomes research: are clinical data derived from administrative databases a reliable alternative to chart review? *Journal of Clinical Epidemiology.* 2000; 53(4): 343-349.

41. Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika.* 1999; 86(4): 843-855.

42. Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology.* 1990; 132(4): 734-745.

43. Spiegelman D, Rosner B, Logan R. Estimation and inference for logistic regression with covariates misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association.* 2000; 95(449): 51-61.

44. Huang Y, Wang CY. Consistent functional methods for logistic regression with errors in covariates. *Journal of the American Statistical Association.* 2001; 96(456): 1469-1482.

45. Spiegelman D, Carroll RJ, Kipnis V. Efficient regression calibration for logistic regression in main study/internal validation study designs with an imperfect reference instrument. *Statistics in Medicine.* 2001; 20(1): 139-160.

46. Rabe-Hesketh S, Pickles A, Skrondal A. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling.* 2003; 3(3): 215.

47. Hossain S, Gustafson P. Bayesian adjustment for covariate measurement errors: a flexible parametric approach. *Statistics in Medicine.* 2009; 28(11): 1580-1600.

48. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine.* 1989; 8(9): 1051-1069; discussion 1071-1053.

49. Carroll RJ, Stefanski LA. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association.* 1990; 85(411): 652-663.

50. Thoresen M. Correction for measurement error in multiple logistic regression: A simulation study. *Journal of Statistical Computation and Simulation.* 2006; 76(6): 475-487.

51. Dalen I, Buonaccorsi JP, Laake P, Hjartaker A, Thoresen M. Regression analysis with categorized regression calibrated exposure: some interesting findings. *Emerging Themes in Epidemiology.* 2006; 3: 6.

52. Satten GA, Kupper LL. Inferences about exposure-disease associations using probability-of- exposure information. *Journal of the American Statistical Association.* 1993; 88(421): 200-208.

53. Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine.* 1997; 16(1-3): 103-116.

54. Spiegelman D, Casella M. Fully parametric and semi-parametric regression models for common events with covariate measurement error in main study/validation study designs. *Biometrics.* 1997; 53(2): 395-409.

55. Higdon R, Schafer DW. Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis.* 2001; 35(3): 283-299.

56. Pepe MS. Inference Using surrogate outcome data and a validation sample. *Biometrika.* 1992; 79(2): 355-365.

57. White H. Maximum likelihood estimation of misspecified models. *Econometrica.* 1982; 50(1): 1-25.

58. Messer K, Natarajan L. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine.* 2008; 27(30): 6332-6350.

59. Buonaccorsi JP. Measurement error in the response in the general linear model. *Journal of the American Statistical Association.* 1996; 91(434): 633-642.

60. Yucel RM, Zaslavsky AM. Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association.* 2005; 100(472): 1123-1132.

61. He Y, Yucel R, Zaslavsky AM. Misreporting, missing data, and multiple imputation: improving accuracy of cancer registry databases. *Chance (N Y).* 2008; 21(3): 55-58.

62. Mallick BK, Gelfand AE. Semiparametric errors-in-variables models a Bayesian approach. *Journal of Statistical Planning and Inference.* 1996; 52(3): 307-321.

63. Chakraborty S, Banerjee T. Analysis of mixed outcomes: misclassified binary responses and measurement error in covariates. *Journal of Statistical Computation and Simulation.* 2009; 80(11): 1197-1209.

64. Schafer J, Ezzati-Rice T, Johnson W, Khare M, Little R, Rubin D. The NHANES III mutiple imputation project. *The U.S. National Center for Health Statistics,* technical report, 1996.

65. Kennickell AB. Multiple imputation in the survey of consumer finances. Paper presented at: Proceedings of the Section on Business and Economic Statistics, 1998 Annual Meetings of the American Statistical Association.

66. Schenker N, Raghunathan TE, Chiu P-L, Makuc DM, Zhang G, Cohen AJ. Multiple imputation of missing income data in the national health interview survey. *Journal of the American Statistical Association.* 2006; 101(475): 924-933.

67. He Y, Zaslavsky AM, Landrum MB, Harrington DP, Catalano P. Multiple imputation in a large-scale complex survey: a practical guide. *Statistical Methods in Medical Research.* 2010; 19(6): 653-670.

68. Little RJA, Rubin DB. *Statistical analysis with missing data*. Hoboken: Wiley; 2002.

69. Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods.* 2001; 6(4): 330-351.

70. Zhou XH, Eckert GJ, Tierney WM. Multiple imputation in public health research. *Statistics in Medicine.* 2001; 20(9-10): 1541-1549.

71. Ayanian JZ, Zaslavsky AM, Fuchs CS, et al. Use of adjuvant chemotherapy and radiation therapy for colorectal cancer in a population-based cohort. Journal of Clinical Oncology. 2003; 21(7): 1293-1300.

72. Schafer JL. *Analysis of Incomplete Multivariate Data.* London: Chapman & Hall; 1997.

73. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology.* 1993; 138(6): 430-442.

74. Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research.* 1999; 8(1): 3-15.

75. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science.* 2007; 8(3): 206-213.

76. Yuan Y. Multiple imputation using SAS software. *Journal of Statistical Software.* 2011; 45(6): 1-25.

77. Juned Siddique OH. MIDAS: A SAS Macro for multiple imputation using distance-aided selection of donors. *Journal of Statistical Software.* 2009; 29(9): 1-18.

78. J. HN, R. LS. Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician.* 2012; 55: 244-254.

79. Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease. Relationships and methods of correction. *Annual Review of Public Health.* 1993; 14: 69-93.

80. Cordeiro GM, McCullagh P. Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological).* 1991; 53(3): 629-643.

81. Clement FM, James MT, Chin R, et al. Validation of a case definition to define chronic dialysis using outpatient administrative data. *BMC Medical Research Methodology.* 2011; 11: 25.

82. Denburg MR, Haynes K, Shults J, Lewis JD, Leonard MB. Validation of The Health Improvement Network (THIN) database for epidemiologic studies of chronic kidney disease. *Pharmacoepidemiol Drug Safety.* 2011; 20(11): 1138-1149.

83. Clogg CC, Rubin DB, Schenker N, Schultz B, Weidman L. Multiple imputation of industry and occupation codes in census public-use samples using Bayesian logistic regression. *Journal of the American Statistical Association.* 1991; 86(413): 68-78.

84. Gilks WR, Richardson S, Spiegelhalter D. *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics.* London: Chapman & Hall; 1996.

85. Egede LE. Effect of comorbid chronic diseases on prevalence and odds of depression in adults with diabetes. *Psychosomatic Medicine.* 2005; 67(1): 46-51.

86. Pan SY, Johnson KC, Ugnat AM, Wen SW, Mao Y. Association of obesity and cancer risk in Canada. *American Journal of Epidemiology.* 2004; 159(3): 259-268.

87. Thompson ML, Myers JE, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? *Occupational and Environmental Medicine.* 1998; 55(4): 272-277.

88. Thoresen M, Laake P. A simulation study of measurement error correction methods in logistic regression. *Biometrics.* 2000; 56(3): 868-872.

89. SAS Institute Inc. *SAS/STAT 9.1 User's Guide.* Cary: SAS Institute Inc.; 2004.

90. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology.* 2001; 54(10): 979-985.

91. Lipsitz SR, Parzen M, Ewell M. Inference using conditional logistic regression with missing covariates. *Biometrics.* 1998; 54(1): 295-303.

92. Raghunathan TEL, J.M. Van Hoewyk, J. Solenberger, P. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology.* 2001; 27(1): 85-95.

93. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis.* Boca Raton: Chapman & Hall/CRC; 2003.

94. Gelman A, Jakulin A, Pittau MG, Su Y-S. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics.* 2008-12 2008; 2(4): 1360-1383.

95. Rubin DB. Multiple Imputation After 18+ Years. *Journal of the American Statistical Association.* 1996; 91(434): 473-489.

APPENDIX A: EXTRA RESULTS FOR SCENARIO 1

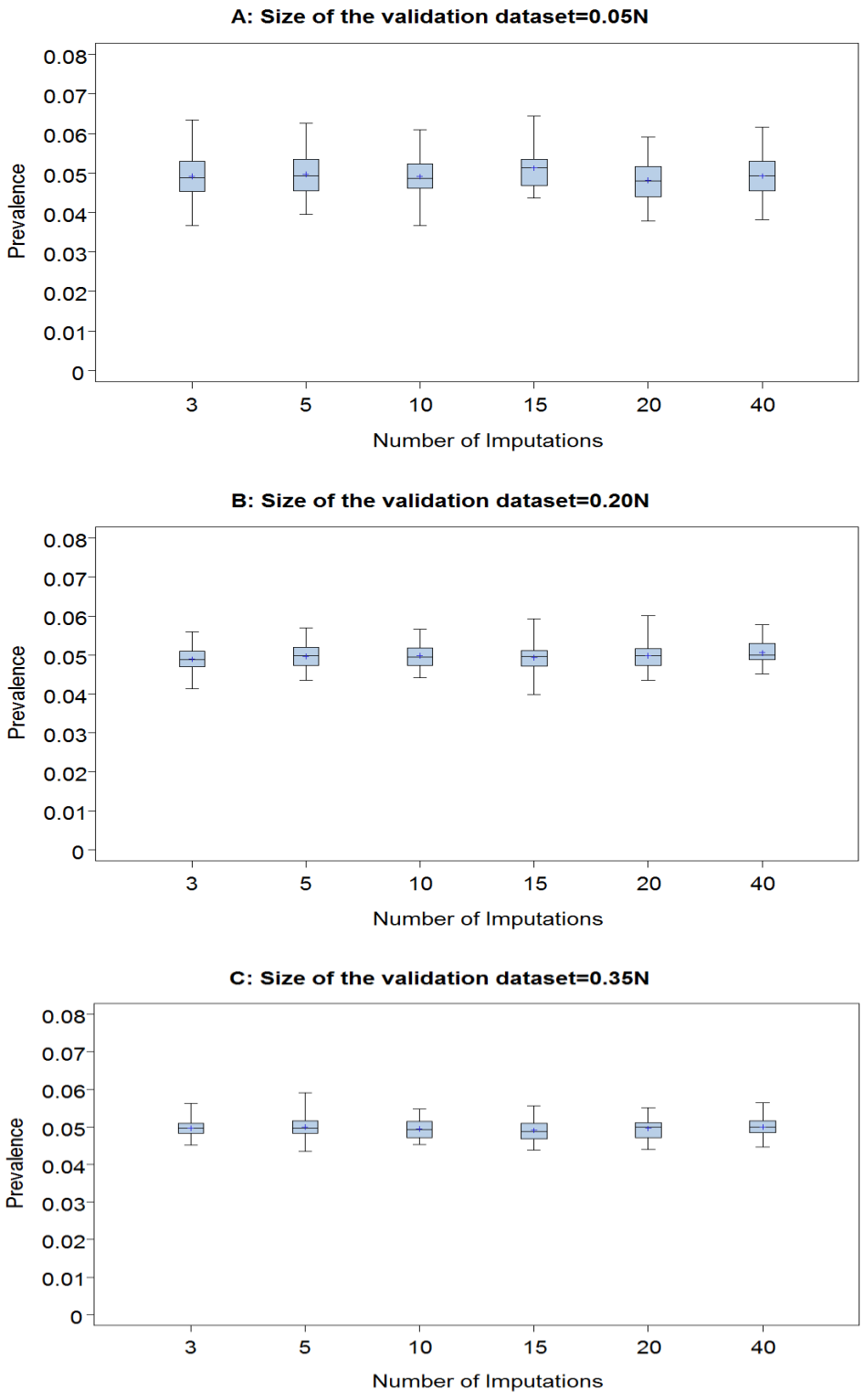

Figure A-1 Prevalence estimates for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.75

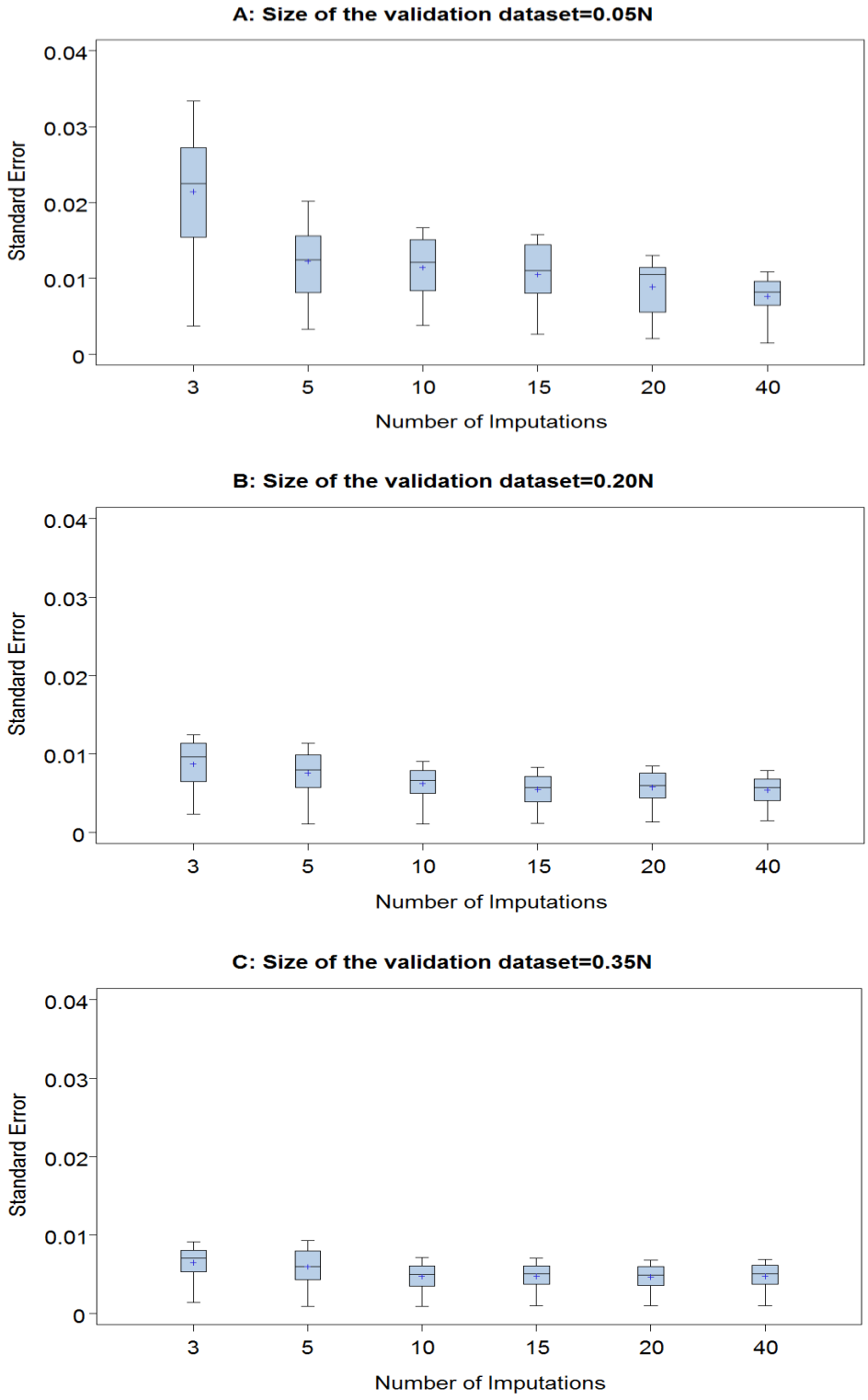

Figure A-2 Standard error of prevalence estimate for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.75

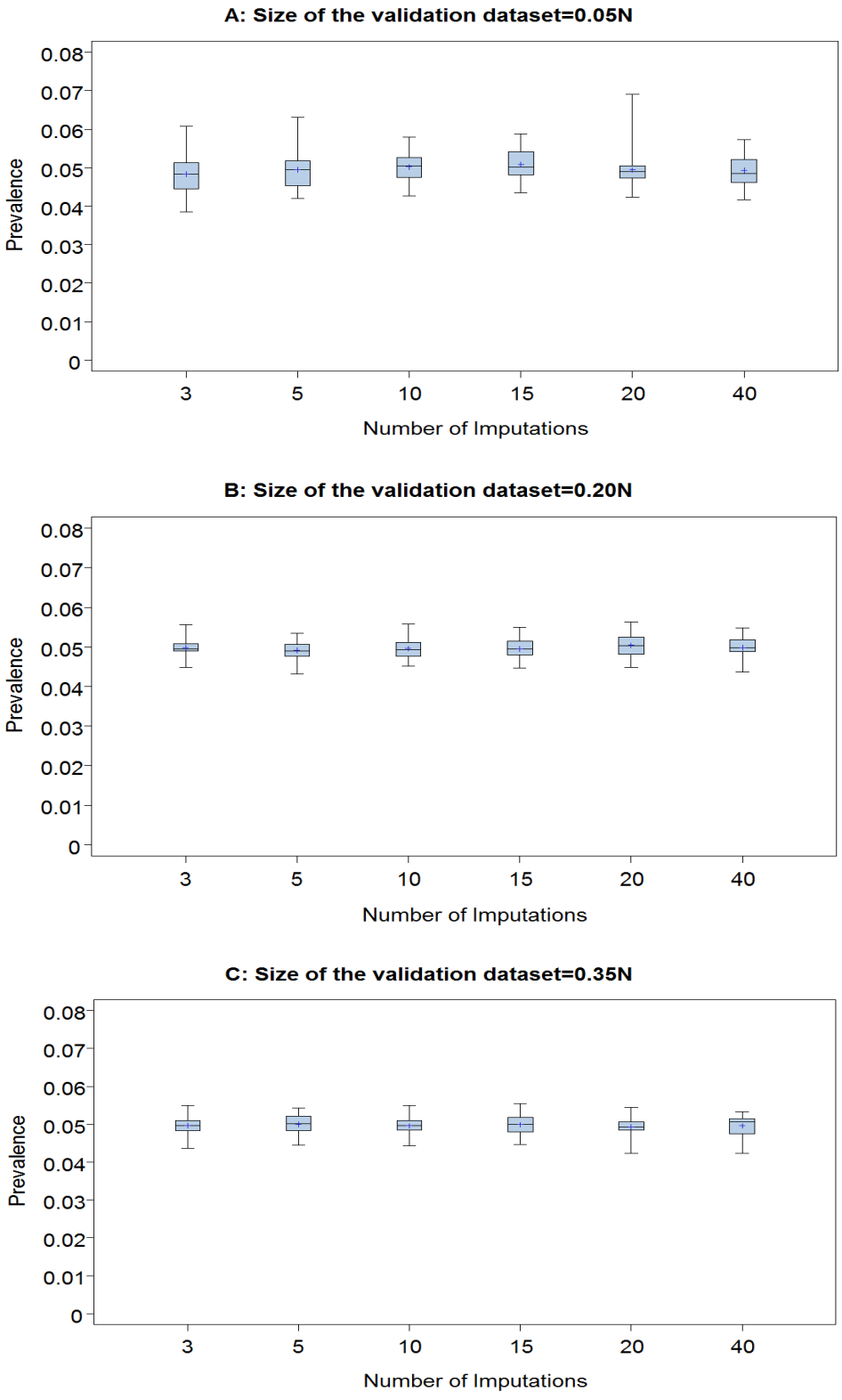

Figure A-3 Prevalence estimates for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.90

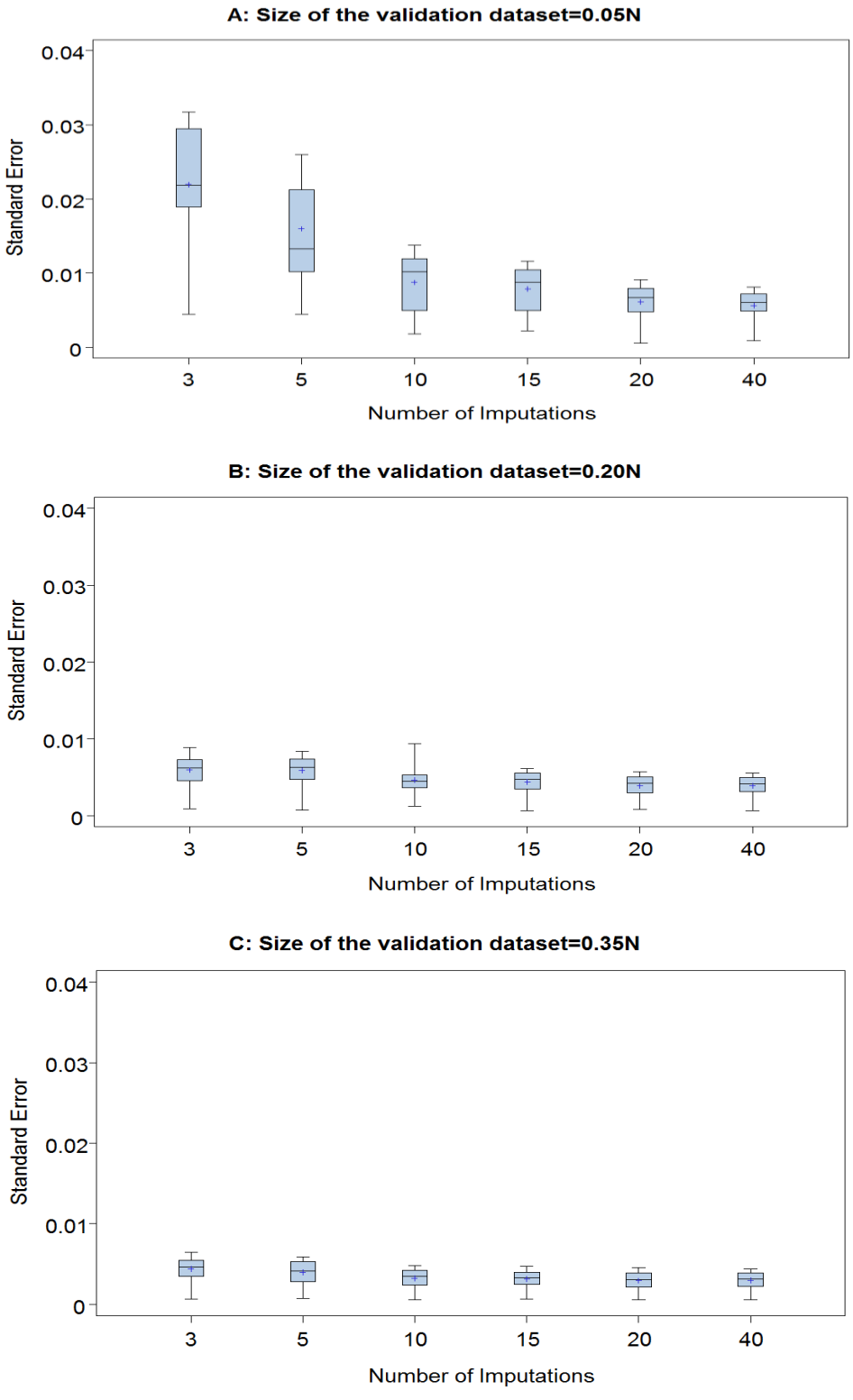

Figure A-4 Standard error of prevalence estimate for the Frequentist MI model with bias correction when the missingness mechanism is MCAR and sensitivity is 0.90

APPENDIX B

Table B-1: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.05N$, Frequentist MI model without bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-20.29	-3.70	-26.69
	MAR	-22.83	-40.14	-27.66
	MNAR	119.68	-37.06	-23.45
0.75	MCAR	11.56	-3.59	-32.07
	MAR	7.23	-48.14	-33.78
	MNAR	269.57	-39.44	55.86
0.90	MCAR	326.42	-3.57	25.50
	MAR	321.73	-39.72	20.64
	MNAR	566.74	-37.88	320.38
			RMSE	
0.60	MCAR	0.0169	0.0048	0.0147
	MAR	0.0167	0.0206	0.0157
	MNAR	0.0790	0.0291	0.0329
0.75	MCAR	0.0276	0.0053	0.0186
	MAR	0.0256	0.0247	0.0193
	MNAR	0.1518	0.0286	0.0640
0.90	MCAR	0.1797	0.0050	0.0467
	MAR	0.1778	0.0271	0.0445
	MNAR	0.2967	0.0285	0.1937
			Coverage	
0.60	MCAR	0.28	0.49	0.33
	MAR	0.28	0.07	0.32
	MNAR	0.13	0.02	0.06
0.75	MCAR	0.27	0.49	0.28
	MAR	0.26	0.03	0.27
	MNAR	0.18	0.02	0.12
0.90	MCAR	0.25	0.49	0.25
	MAR	0.24	0.02	0.24
	MNAR	0.18	0.02	0.21

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table B-2: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.20N$, Frequentist MI model without bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-27.06	-5.28	-23.52
	MAR	-31.22	-42.82	-27.51
	MNAR	-53.52	-66.91	-53.85
0.75	MCAR	-31.93	-5.30	-29.37
	MAR	-36.98	-51.81	-34.22
	MNAR	-37.64	-67.00	-50.78
0.90	MCAR	62.96	-5.27	-35.22
	MAR	50.10	-60.76	-41.24
	MNAR	202.53	-66.94	-27.71
			RMSE	
0.60	MCAR	0.0144	0.0038	0.0127
	MAR	0.0166	0.0214	0.0148
	MNAR	0.0276	0.0333	0.0274
0.75	MCAR	0.0170	0.0038	0.0158
	MAR	0.0197	0.0256	0.0184
	MNAR	0.0238	0.0333	0.0263
0.90	MCAR	0.0481	0.0038	0.0189
	MAR	0.0434	0.0303	0.0221
	MNAR	0.1167	0.0333	0.0285
			Coverage	
0.60	MCAR	0.21	0.63	0.27
	MAR	0.19	0.00	0.26
	MNAR	0.03	0.00	0.04
0.75	MCAR	0.20	0.63	0.24
	MAR	0.19	0.00	0.23
	MNAR	0.08	0.00	0.08
0.90	MCAR	0.22	0.63	0.23
	MAR	0.22	0.00	0.22
	MNAR	0.22	0.00	0.14

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

Table B-3: Relative bias, RMSE and 95% confidence interval coverage when the size of the validation dataset is $0.35N$, Frequentist MI model without bias correction

Sensitivity of observed disease status	Missing mechanism	Model 1	Model 2	Model 3
			Relative Bias (%)	
0.60	MCAR	-22.21	-4.62	-19.15
	MAR	-29.32	-42.30	-26.30
	MNAR	-53.42	-66.94	-52.62
0.75	MCAR	-26.13	-4.53	-23.94
	MAR	-34.83	-51.63	-32.99
	MNAR	-49.10	-66.96	-49.24
0.90	MCAR	-0.31	-4.60	-28.94
	MAR	-13.44	-60.51	-39.24
	MNAR	88.17	-66.89	-44.80
			RMSE	
0.60	MCAR	0.0118	0.0033	0.0104
	MAR	0.0156	0.0211	0.0141
	MNAR	0.0271	0.0333	0.0267
0.75	MCAR	0.0139	0.0033	0.0129
	MAR	0.0185	0.0257	0.0176
	MNAR	0.0254	0.0333	0.0253
0.90	MCAR	0.0178	0.0033	0.0155
	MAR	0.0191	0.0301	0.0210
	MNAR	0.0586	0.0333	0.0241
			Coverage	
0.60	MCAR	0.19	0.73	0.27
	MAR	0.19	0.00	0.24
	MNAR	0.03	0.00	0.03
0.75	MCAR	0.20	0.74	0.23
	MAR	0.20	0.00	0.22
	MNAR	0.06	0.00	0.07
0.90	MCAR	0.22	0.74	0.22
	MAR	0.22	0.00	0.22
	MNAR	0.19	0.00	0.13

Note: Model 1 is the predictive model with the observed disease status and the imputed disease predictors as covariates; Model 2 is the predictive model with the imputed disease predictors as covariates only; Model 3 is the predictive model with the observed disease status as covariate only. RMSE = root mean squared error. MCAR denotes the missing completely at random mechanism; MAR denotes the missing at random mechanism; MNAR denotes the missing not at random mechanism.

APPENDIX C

Based on the conditional distribution, the predictive model for imputing the missing true response in the main dataset is conditional on the information of observed response, true covariates and observed covariates.

$$P(Y|U,X_1,X_2,W_1,W_2) = \frac{P(Y,U,X_1,X_2,W_1,W_2)}{P(U,X_1,X_2,W_1,W_2)} \qquad \text{(C-1)}$$

To factor the joint distribution $P(Y,U,X_1,X_2,W_1,W_2)$ is dependent on the assumptions of the variables. The true disease predictors X_1 and X_2 are independent. From section 3.2, the forms of the measurement error models are clarified. When the misclassification of the observed response is independent on the disease predictors which is in Scenario 1 of our study, the decomposition of the joint distribution of $P(Y,U,X_1,X_2,W_1,W_2)$ is

$$P(Y,U,X_1,X_2,W_1,W_2) = P(U|Y)P(Y|X_1,X_2)P(X_1|W_1)P(X_2|W_2)P(W_1)P(W_2) \qquad \text{(C-2)}$$

The relationship between the observed disease status U and the true disease status Y is described by measures of sensitivity and specificity. The probability of Y conditional on X_1 and X_2 is the disease model in Equation 1, which reflects the relationship between the disease and predictors. When the misclassification of the observed disease status is dependent on the disease predictors, the decomposition of the joint distribution of $P(Y,U,X_1,X_2,W_1,W_2)$ is

$$P(Y,U,X_1,X_2,W_1,W_2) = P(U|Y,X_1,X_2)P(Y|X_1,X_2)P(X_1|W_1)P(X_2|W_2)P(W_1)P(W_2) \text{ (C-3)}$$

The relationship between the observed disease status U and the true disease status Y is dependent on the true disease predictors, which means the sensitivity and specificity of the observed disease status U varies across different values of the disease predictors. It characterizes how the errors appear in the observed response. Also, the probability of Y conditional on X_1 and X_2 is the disease model as Equation 1 that reflects the relationship between the disease and disease markers.

We develop the sequential regression MI method first for the scenario 1 when the misclassification of the observed disease status is independent of the disease predictors. Based on the decomposition of the joint distribution of $P(Y,U,X_1,X_2,W_1,W_2)$ as Equation 23, we can have

$$P(Y = 1|U, X_1, X_2, W_1, W_2)$$

$$= \frac{P(U = u|Y = 1)f_1(Y = 1|x_1, x_2)f(x_1|w_1)f(x_2|w_2)f(w_1)f(w_2)}{\sum_{y=0,1} P(Y = y, U = u, X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2)}$$

$$= \frac{P(U = u|Y = 1)f_1(Y = 1|x_1, x_2)}{P(U = u|Y = 1)f_1(Y = 1|x_1, x_2) + P(U = u|Y = 0)(1 - f_1(Y = 1|x_1, x_2))}$$

As for the sequential regression MI method for the scenario 2 when the misclassification of the observed disease status is dependent on the disease predictors, based on the decomposition of the joint distribution of $P(Y, U, X_1, X_2, W_1, W_2)$ as Equation 24 we can have

$$P(Y = 1|U, X_1, X_2, W_1, W_2)$$

$$= \frac{P(U = u|Y = 1, x_1, x_2)f_1(Y = 1|x_1, x_2)f(x_1|w_1)f(x_2|w_2)f(w_1)f(w_2)}{\sum_{y=0,1} P(Y = y, U = u, X_1 = x_1, X_2 = x_2, W_1 = w_1, W_2 = w_2)}$$

Based on this predictive model for true disease status, the probability of disease conditional on the disease predictors need to be estimated from validation dataset, the sensitivity and specificity of the observed disease status also need to be estimated from the validation dataset. Since in the main dataset the true values of the covariates are unobservable, this predictive model also needs more information to be applied to correct for measurement error.

APPENDIX D

Computer Simulation Program for Scenario 1: Misclassification of Observed Disease Status is Independent of Disease Predictors

Following program code were run for the conditions that when the sensitivity of the observed disease status was 0.60, the size of the validation dataset was 0.05N, the number of imputations including 3, 5, 10, 15, 20 and 40, and the missingness mechanism was specified as MCAR, MAR or MNAR.

```
Libname MCAR 'C:\MISIM Program\MCAR';
proc iml;
** Specify simulation parameters**;
ntot=10000; **total number of observations in the dataset**;
specificity=1.0; **Specificity of whole population dataset**;
nsim=500; **number of simulations**;
sensitivity=0.6;
valrate=0.05;
numberofimpute={3 5 10 15 20 40};
beta0=-2.785; **Specify the coefficient value**;
beta1=0.539;
beta2=-0.693;
**create counters for disease state**;
indtrue=j(ntot,1,.); **Temporary indicator of true disease status**;
indobs=j(ntot,1,.); **Temporary indicator of observed disease status**;
indval=j(ntot,1,.); **Temporary indicator of presence on the validation cohort**;
ind=j(ntot,1,.);
alldata=j(ntot,3,.);
**ALLDATA:
```

```
first column = member of the validation cohort (presence=1, otherwise=0),
second column = true disease state (presence=1, otherwise=0),
third column = observed disease status (presence=1, otherwise=0);

start binest(beta,cov,y,x,b); **Newton-Raphson Algorithm**;
        a=nrow(x);
        b=repeat(0,ncol(x),1);
        p=j(nrow(x),1,.);
        oldb=b+1;
        do iter=1 to 20 while(max(abs(b-oldb))>1e-8);
                oldb=b;
                z=x*b;
                do j=1 to a;
                        if z[j]>0 then pj=1/(1+exp(-z[j]));
                        else pj=exp(z[j])/(1+exp(z[j]));
                        p[j]=pj;
                end;
                g=x`*(y-p);
                w=diag(p*(1-p)`);
                f=x`*w*x;
                b=b+ginv(f)*g;
        end;
        beta=b;
        cov=ginv(f);
finish;

**Bias correction by Cordeiro and McCullagh**;
```

```
start biascor(bc_beta,bc_cov,y,x,b);
        a=nrow(x);
        b=repeat(0,ncol(x),1);
        p=j(nrow(x),1,.);
        oldb=b+1;
        do iter=1 to 20 while(max(abs(b-oldb))>1e-8);
                oldb=b;
                z=x*b;
                do j=1 to a;
                        if z[j]>0 then pj=1/(1+exp(-z[j]));
                        else pj=exp(z[j])/(1+exp(z[j]));
                        p[j]=pj;
                end;
                w=diag(p*(1-p)`);
                f=x`*w*x;
                d=ginv(f);
                g=x`*(y-p);
                e=diag(x*d*x`)*(p-0.5);
                c=d*x`*w*e;
                b=b+d*g-c;
        end;
        bc_beta=b;
        bc_cov=d;
finish;

**Build model to generate vector of regression coefficients for multiple imputations**;
start betagen(tx,beta,cov,nimpute,seed);
```

```
        v=nrow(cov);
        do i=1 to v; **Make sure the covariance matrix is symmetric**;
                do j=1 to v;
                        if cov[j,i]^=cov[i,j] then cov[j,i]=cov[i,j];
                end;
        end;
        l=t(root(cov));
        z=normal(j(v,nimpute,seed));
        x=l*z;
        x=repeat(beta,1,nimpute)+x;
        tx=t(x);
finish;


**Specify starting seeds for all random number generators**;
randseed=j(11,1,.);
do i=1 to 11;
        randseed[i]=int(10000*ranuni(0)); **the seed 0 make sure the ranuni generate differently
random number each time**;
end;
seed1=randseed[1];
seed2=randseed[2];
seed3=randseed[3];
seed4=randseed[4];
seed5=randseed[5];
seed6=randseed[6];
seed7=randseed[7];
seed8=randseed[8];
seed9=randseed[9];
```

```
seed10=randseed[10];

seed11=randseed[11];


ResultAll=j(6,29,.);


do ni=1 to 6;

        nimpute=numberofimpute[ni];

yimpute=j(ntot,nimpute,.); **YIMPUTE: values of disease status for imputation based on logistic regression**;

yimpute_bc=j(ntot,nimpute,.); **YIMPUTE_BC: values of disease status for imputation based on bias correction logistic**;

b_yimpute=j(ntot,nimpute,.);**BYIMPUTE: values of disease status for imputaton based on bayesian method**;

prev=0;          **true value of prevalence**;

obs_prev=0; **value of observed prevalence**;

obs_bias=0; **The bias of the observed prevalence**;

SN_U=0; **The sensitivity of the observed response (diagnosis U)**;

SP_U=0;

mim_prev=0; **value of prevalence based on bias correction logistic with multiple imputation**;

bias=0;

rel_bias=0;

st_bias=0;

tvar=0;**Total variability of bias correction MIM based on observed response**;

coverage=0;

bcmim_prev=0; **value of prevalence based on bias correction logistic with multiple imputation**;

bc_bias=0;

bcrel_bias=0;
```

```
bcst_bias=0;
bc_tvar=0;**Total variability of bias correction MIM based on observed response**;
bc_coverage=0;
B_prev=0; **value of prevalence based in Bayesian approach**;
B_bias=0; **value of bias in Bayesian approach**;
B_relbias=0;
B_stbias=0;
B_tvar=0;
B_coverage=0;

**Main Body of Simulation**;
do r=1 to nsim;
**Generate a continuous covariate data and a binary covariate data**;
x1=j(ntot,1,.);
call randseed(seed1);
call randgen (x1,'normal',0,1);
x2=j(ntot,1,.);
call randseed(seed2);
call randgen (x2,'Bernoulli',.50);

**Create true disease status**;
do i=1 to ntot;
        pi=exp(beta0+beta1*x1[i]+beta2*x2[i])/(1+exp(beta0+beta1*x1[i]+beta2*x2[i]));
        call randseed(seed3);
        call randgen(yi,'bernoulli',pi);
        alldata[i,2]=yi;
end;
```

```
prev1=sum(alldata[,2])/ntot;
ytrue=alldata[,2];
prev=prev+prev1;

**Create observed disease status**;
call randseed(seed4);
call randgen(indobs,'bernoulli',sensitivity); **Depend on Sensitivity only**;
do i=1 to ntot;
        if alldata[i,2]=1 then alldata[i,3]=indobs[i];
        else alldata[i,3]=0;
end;

obsprev1=sum(alldata[,3])/ntot;
obs_prev=obs_prev+obsprev1;
obs_bias1=prev1-obsprev1;
obs_bias=obs_bias+obs_bias1;
**Estimate the sensitivity and specificity of the diagnosis(U)**;
SN_count=0;
SP_count=0;
do i=1 to ntot;
        if alldata[i,2]=1 & alldata[i,3]=1 then SN_count=SN_count+1;
        else if alldata[i,2]=0 & alldata[i,3]=0 then SP_count=SP_count+1;
end;
SN_U=SN_U+SN_count/sum(alldata[,2]);
SP_U=SP_U+SP_count/(ntot-sum(alldata[,2]));

**Create indicator for membership in the validation cohort and output validation cohort data into
a new dataset for analysis**;
```

```
**MCAR**;
call randseed(seed5);
call randgen(indval,'uniform');
validss=valrate*ntot;
temp=rank(indval);
yset=j(validss,5,1);
count=0;
do i=1 to ntot;
        if temp[i]>validss then alldata[i,1]=0;
        else if temp[i]<=validss then do;
                alldata[i,1]=1;
                count=count+1;
                yset[count,1]=alldata[i,2];
                yset[count,2]=1;
                yset[count,3]=alldata[i,3];
                yset[count,4]=x1[i];
                yset[count,5]=x2[i];
        end;
end;
Varname={y n u x1 x2};
create validation from yset [colname=varname];
append from yset;
close validation;
Val= "Work.Validation";

**Original estimates of logistic with observed response**;
call binest(beta,cov,yset[,1],yset[,2:3],b);
```

```
call betagen(betamat,beta,cov,nimpute,seed6);
storval=j(nimpute,1,.);
storvar=j(nimpute,1,.);

do k=1 to nimpute;
	prob=j(ntot,1,.);
	fill_vec=j(ntot,1,.);
	indvar=j(ntot,1,.);
	do i=1 to ntot;
		if betamat[k,1]+betamat[k,2]*alldata[i,3]>0 then p=1/(1+exp(-
(betamat[k,1]+betamat[k,2]*alldata[i,3])));
		else
p=exp(betamat[k,1]+betamat[k,2]*alldata[i,3])/(1+exp(betamat[k,1]+betamat[k,2]*alldata[i,3]));
		prob[i]=p;
		if alldata[i,1]=0 then do;
			call randseed(seed7);
			call randgen(fill,'bernoulli',prob[i]);
			fill_vec[i]=fill;
		end;
		else if alldata[i,1]=1 then fill_vec[i]=alldata[i,2];
	end;
	yimpute[,k]=fill_vec;
	fprev=sum(fill_vec)/ntot;
	do i=1 to ntot;
		indvar[i]=prob[i]*(1-prob[i]);
	end;
	predvar=sum(indvar)/ntot**2;
	storval[k]=fprev;
```

```
        storvar[k]=predvar;
end;
fprev0=sum(storval)/nimpute;
mim_prev=mim_prev+fprev0;
bias1=fprev0-prev1;
bias=bias+bias1;
rel_bias1=bias1/prev1; **The bias relative to the true prevelance**;
rel_bias=rel_bias+rel_bias1;
wvar=sum(storvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM based on
observed response only**;

do q=1 to nimpute;
        bvar=sum((storval[q]-sum(storval)/nimpute)**2)/(nimpute-1);
end;

tvar0=(wvar+(1+1/nimpute)*bvar); **Total variability of Bias Correction MIM based on
observed response only**;
tvar=tvar+tvar0;
SE=sqrt(tvar0);
st_bias1=bias1/SE;
st_bias=st_bias+st_bias1;
LCI=prev-1.96*SE;
UCI=prev+1.96*SE;
a=LCI<=prev1 & prev1<=UCI;
coverage=coverage+a;

**Bias Correction estimates of logistic with observed response**;
call biascor(bc_beta,bc_cov,yset[,1],yset[,2:3],b);
```

```
call betagen(bcbetamat,bc_beta,bc_cov,nimpute,seed8);
bc_storval=j(nimpute,1,.);
bc_storvar=j(nimpute,1,.);


do k=1 to nimpute;
	bc_prob=j(ntot,1,.);
	bcfill_vec=j(ntot,1,.);
	bc_indvar=j(ntot,1,.);
	do i=1 to ntot;
		if bcbetamat[k,1]+bcbetamat[k,2]*alldata[i,3]>0 then bcp=1/(1+exp(-
(bcbetamat[k,1]+bcbetamat[k,2]*alldata[i,3])));
		else
bcp=exp(bcbetamat[k,1]+bcbetamat[k,2]*alldata[i,3])/(1+exp(bcbetamat[k,1]+bcbetamat[k,2]*al
ldata[i,3]));
		bc_prob[i]=bcp;
		if alldata[i,1]=0 then do;
			call randseed(seed9);
			call randgen(bcfill,'bernoulli',bc_prob[i]);
			bcfill_vec[i]=bcfill;
		end;
		else if alldata[i,1]=1 then bcfill_vec[i]=alldata[i,2];
	end;
	yimpute_bc[,k]=bcfill_vec;
	bcprev=sum(bcfill_vec)/ntot;
	do i=1 to ntot;
		bc_indvar[i]=bc_prob[i]*(1-bc_prob[i]);
	end;
	bcpredvar=sum(bc_indvar)/ntot**2;
```

```
        bc_storval[k]=bcprev;
        bc_storvar[k]=bcpredvar;
end;

bc_prev=sum(bc_storval)/nimpute;
bcmim_prev=bcmim_prev+bc_prev;
bc_bias1=bc_prev-prev1;
bc_bias=bc_bias+bc_bias1;
bcrel_bias1=bc_bias1/prev1; **The bias relative to the true prevelance**;
bcrel_bias=bcrel_bias+bcrel_bias1;
bc_wvar=sum(bc_storvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM based on observed response only**;

do q=1 to nimpute;
        bc_bvar=sum((bc_storval[q]-sum(bc_storval)/nimpute)**2)/(nimpute-1);
end;

bc_tvar0=(bc_wvar+(1+1/nimpute)*bc_bvar); **Total variability of Bias Correction MIM based on observed response only**;
bc_tvar=bc_tvar+bc_tvar0;
bc_SE=sqrt(bc_tvar0);
bcst_bias1=bc_bias1/bc_SE;
bcst_bias=bcst_bias+bcst_bias1;
bc_LCI=bc_prev-1.96*bc_SE;
bc_UCI=bc_prev+1.96*bc_SE;
bca=bc_LCI<=prev1 & prev1<=bc_UCI;
bc_coverage=bc_coverage+bca;
```

```
submit Val;
proc mcmc data=&Val ntu=1000 nmc=21000 nthin=100 nbi=1000 outpost=valest seed=2481;
        ods select PostSummaries PostIntervals mcse ess ;
        parms (alpha0 alpha1) 0;
        prior alpha0 alpha1 ~ normal(0, var=1000);
        p = logistic(alpha0+alpha1*u);
        model Y ~ binomial(n,p);
run;
endsubmit;

declare DataObject dobj;
dobj=DataObject.CreateFromServerDataSet("Work.valest");
postn=dobj.GetNumObs();
dobj.GetVarData("alpha0",alpha0);
dobj.GetVarData("alpha1",alpha1);

postsample=alpha0||alpha1;
randind=j(nimpute,1,.);
call randseed(seed10);
call randgen(randind,'uniform');
MIsample=j(nimpute,2,.);
B_Yimpute=j(ntot,nimpute,.);
B_storval=j(nimpute,1,.);
B_storvar=j(nimpute,1,.);

do j=1 to nimpute;
```

```
        B_prob=j(ntot,1,.);
        B_fill_vec=j(ntot,1,.);
        B_indvar=j(ntot,1,.);
        MIind=int(randind[j,]*postn)+1;
        MIsample[j,]=postsample[MIind,];
        do i=1 to ntot;
                if MIsample[j,1]+MIsample[j,2]*alldata[i,3]>0 then bp=1/(1+exp(-
(MIsample[j,1]+MIsample[j,2]*alldata[i,3])));
                else bp=exp(MIsample[j,1]+MIsample[j,2]*alldata[i,3])/(1+exp(-
(MIsample[j,1]+MIsample[j,2]*alldata[i,3])));
                B_prob[i]=bp;
                if alldata[i,1]=0 then do;
                        call randseed(seed11);
                        call randgen(B_fill,'bernoulli',B_prob[i]);
                        B_fill_vec[i]=B_fill;
                end;
                else if alldata[i,1]=1 then B_fill_vec[i]=alldata[i,2];
        end;
        B_Yimpute[,j]=B_fill_vec;
        Bprev=sum(B_Yimpute[,j])/ntot;
        do i=1 to ntot;
                B_indvar[i]=B_prob[i]*(1-B_prob[i]);
        end;
        Bpredvar=sum(B_indvar)/ntot**2;
        B_storval[j]=Bprev;
        B_storvar[j]=Bpredvar;
end;
```

```
B_prev0=sum(B_storval)/nimpute;

B_prev=B_prev+B_prev0;

B_bias0=B_prev0-prev1;

B_bias=B_bias+B_bias0;

B_relbias0=B_bias0/prev1; **The bias relative to the true prevelance**;

B_relbias=B_relbias+B_relbias0;

B_wvar=sum(B_storvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM
based on observed response only**;


do q=1 to nimpute;

        B_bvar=sum((B_storval[q]-sum(B_storval)/nimpute)**2)/(nimpute-1);

end;


B_tvar0=(B_wvar+(1+1/nimpute)*B_bvar); **Total variability of Bias Correction MIM based
on observed response only**;

B_tvar=B_tvar+B_tvar0;

B_SE=sqrt(B_tvar0);

B_stbias0=B_bias0/B_SE;

B_stbias=B_stbias+B_stbias0;

B_LCI=B_prev0-1.96*B_SE;

B_UCI=B_prev0+1.96*B_SE;

Ba=B_LCI<=prev1 & prev1<=B_UCI;

B_coverage=B_coverage+Ba;

end;


prev=prev/nsim;

obs_prev=obs_prev/nsim;

obs_bias=obs_bias/nsim;
```

```
SN_U=SN_U/nsim;
SP_U=SP_U/nsim;
mim_prev=mim_prev/nsim;
bias=bias/nsim;
rel_bias=rel_bias/nsim;
st_bias=st_bias/nsim;
tvar=tvar/nsim;
mse=tvar+bias**2;
rmse=sqrt(mse);
coverage=coverage/nsim;
bcmim_prev=bcmim_prev/nsim;
bc_bias=bc_bias/nsim;
bcrel_bias=bcrel_bias/nsim;
bcst_bias=bcst_bias/nsim;
bc_tvar=bc_tvar/nsim;
bc_mse=bc_tvar+bc_bias**2;
bc_rmse=sqrt(bc_mse);
bc_coverage=bc_coverage/nsim;
B_prev=B_prev/nsim;
B_bias=B_bias/nsim;
B_relbias=B_relbias/nsim;
B_stbias=B_stbias/nsim;
B_tvar=B_tvar/nsim;
B_mse=B_tvar+B_bias**2;
B_rmse=sqrt(B_mse);
B_coverage=B_coverage/nsim;
```

```
SEED=seed1||seed2||seed3||seed4||seed5||seed6||seed7||seed8||seed9||seed10||seed11;

Parameter=ntot||sensitivity||valrate||nimpute;

OBS=nsim||prev||obs_prev||obs_bias;

MIM=mim_prev||bias||rel_bias||st_bias||tvar||rmse||coverage;

BCMIM=bcmim_prev||bc_bias||bcrel_bias||bcst_bias||bc_tvar||bc_rmse||bc_coverage;

B_MIM=B_prev||B_bias||B_relbias||B_stbias||B_tvar||B_rmse||B_coverage;

Result=Parameter||OBS||MIM||BCMIM||B_MIM;

ResultAll[1#(nimpute=3)+2#(nimpute=5)+3#(nimpute=10)+4#(nimpute=15)+5#(nimpute=20)+
6#(nimpute=40),]=Result;


Varnames={TotalNumber Sensitivity ValRate NImpute SimulationNumber Prevalence
ObservedPrevalence ObserPrevBias FIPrev FIBias FIRelBias FIStBias FIVar FIRMSE
FICoverage BCPrev BCBias BCRelBias BCStBias BCVar BCRMSE BCCoverage BIPrev
BIBias BIRelBias BIStBias BIVar BIRMSE BICoverage};

end;

Create MCAR.MCAR05S65 from ResultAll [colname=varnames];

append from ResultAll;

close MCAR.MCAR05S65;

Create MCAR.SEED05S65 from SEED;

append from SEED;

close MCAR.SEED05S65;

quit;
```

Computer Simulation Program for Scenario 2: Misclassification of Observed Disease Status is Dependent on Disease Predictors

Following program code were run for the conditions that when the sensitivity of the observed disease status was specified as 0.60, 0.75 or 0.90, the size of the validation dataset was 0.05N, the number of imputations including 3, 5, 10, 15, 20 and 40, the sensitivity and specificity of the binary disease predictor were 0.90 or 0.70, and the additive variance of the continuous disease predictor was 1 or 2, and the missingness mechanism was specified as MCAR, MAR or MNAR.

```
LIBNAME MCAR2 'C:\MISIM Program\MCAR2';
proc iml;
**Multiple Imputation in the Frequentist Approach**;
**This program is to compare Naive Logistic Regression and Bias Corrected Logistic
Regression for observed response is depend on covariates**;

** Specify simulation parameters**;
ntot=10000; **total number of observations in the dataset**;
nsim=500; **number of simulations**;
sensitivity=0.6;
valrate=0.05;
numberofimpute={3 5 10 15 20 40};
beta0=-2.785; **Specify the coefficient value**;
beta1=0.539;
beta2=-0.693;
**The coefficient for sensitivity 0.60 of the observed disease status**;
rho0=1.2;
rho1=-0.26934;
rho2=-1.7536;
**The coefficient for sensitivity 0.75 of the observed disease status**;
rho0=1.8295;
rho1=-0.269;
rho2=-1.386;
**The coefficient for sensitivity 0.90 of the observed disease status**;
rho0=1.995;
rho1=-0.26934;
rho2=1.36;
**create counters for disease state**;
indval=j(ntot,1,.); **Temporary indicator of presence on the validation cohort**;
alldata=j(ntot,3,.);
indobs1=j(ntot,1,.); ** Temporary indicator of measurment of x1**;
indobs2=j(ntot,1,.);
**ALLDATA:
```

```
first column = member of the validation cohort (presence=1, otherwise=0),
second column = true disease state (presence=1, otherwise=0),
third column = observed disease status (presence=1, otherwise=0);

**Build Logistic Regression model for estimating the parameters**;
start binest(beta,cov,y,x,b); **Newton-Raphson Algorithm**;
        a=nrow(x);
        b=repeat(0,ncol(x),1);
        p=j(nrow(x),1,.);
        oldb=b+1;
        do iter=1 to 20 while(max(abs(b-oldb))>1e-8);
                oldb=b;
                z=x*b;
                do j=1 to a;
                        if z[j]>0 then pj=1/(1+exp(-z[j]));
                        else pj=exp(z[j])/(1+exp(z[j]));
                        p[j]=pj;
                end;
                g=x`*(y-p);
                w=diag(p*(1-p)`);
                f=x`*w*x;
                b=b+ginv(f)*g;
        end;
        beta=b;
        cov=ginv(f);
finish;

**Bias correction by Cordeiro and McCullagh**;
start biascor(bc_beta,bc_cov,y,x,b);
        a=nrow(x);
        b=repeat(0,ncol(x),1);
        p=j(nrow(x),1,.);
        oldb=b+1;
        do iter=1 to 20 while(max(abs(b-oldb))>1e-8);
                oldb=b;
                z=x*b;
                do j=1 to a;
                        if z[j]>0 then pj=1/(1+exp(-z[j]));
                        else pj=exp(z[j])/(1+exp(z[j]));
                        p[j]=pj;
                end;
                w=diag(p*(1-p)`);
                f=x`*w*x;
                d=ginv(f);
                g=x`*(y-p);
                e=diag(x*d*x`)*(p-0.5);
```

```
                c=d*x`*w*e;
                b=b+d*g-c;
        end;
        bc_beta=b;
        bc_cov=d;
finish;

**Linear regression analysis to estimate the parameters of the linear model to impute the true
continuous covariate**;
start regress(y,x,b,covb);
        xpxi=ginv(t(x)*x);
        beta=xpxi*(t(x)*y);
        yhat=x*beta;
        resid=y-yhat;
        sse=ssq(resid);
        dfe=nrow(x)-ncol(x);
        mse=sse/dfe;
        covbeta=xpxi*mse;
        b=beta;
        covb=covbeta;
finish;

**Build model to generate vector of regression coefficients for multiple imputations**;
start betagen(tx,beta,cov,nimpute,seed);
        v=nrow(cov);
        do i=1 to v; **Make sure the covariance matrix is symmetric**;
                do j=1 to v;
                        if cov[j,i]^=cov[i,j] then cov[i,j]=cov[j,i];
                end;
        end;
        l=t(root(cov));
        z=normal(j(v,nimpute,seed));
        x=l*z;
        x=repeat(beta,1,nimpute)+x;
        tx=t(x);
finish;

**Specify starting seeds for all random number generators**;
randseed=j(27,1,.);
do i=1 to 27;
        randseed[i]=int(10000*ranuni(0)); **the seed 0 make sure the ranuni generate differently
random number each time**;
end;
seed1=randseed[1];
seed2=randseed[2];
seed3=randseed[3];
```

```
seed4=randseed[4];
seed5=randseed[5];
seed6=randseed[6];
seed7=randseed[7];
seed8=randseed[8];
seed9=randseed[9];
seed10=randseed[10];
seed11=randseed[11];
seed12=randseed[12];
seed13=randseed[13];
seed14=randseed[14];
seed15=randseed[15];
seed16=randseed[16];
seed17=randseed[17];
seed18=randseed[18];
seed19=randseed[19];
seed20=randseed[20];
seed21=randseed[21];
seed22=randseed[22];
seed23=randseed[23];
seed24=randseed[24];
seed25=randseed[25];
seed26=randseed[26];
seed27=randseed[27];

ResultAll=j(24,69,.);
**Vary the parameters**;
do MEvar=1 to 2; **The varianace of additional measurement error of the continuous covariate
x1**;

        do XSN=1 to 3 by 2;  **Specify the sensitivity and specificity of the covariate's
misclassification**;
                SN=1-XSN/10;
                if SN=0.9 then FP=0.1; **False Positive equals 1-Specificity**;
                if SN=0.7 then FP=0.3;

                        do ni=1 to 6;
                                nimpute=numberofimpute[ni];

prev=0;         **true value of prevalence**;
obs_prev=0; **value of observed prevalence**;
obs_bias=0;**the bias of the observed prevalence**;
SN_U=0; **The sensitivity of the observed response (diagnosis U)**;
SP_U=0; **The specificity of the observed response (diagnosis U)**;

t_prev=0;
```

t_bias=**0**;
t_relbias=**0**;
t_stbias=**0**;
t_tvar=**0**;
t_coverage=**0**;

t_bcprev=**0**;
t_bcbias=**0**;
t_bcrelbias=**0**;
t_bcstbias=**0**;
t_bctvar=**0**;
t_bccoverage=**0**;

o_prev=**0**;
o_bias=**0**;
o_relbias=**0**;
o_stbias=**0**;
o_tvar=**0**;
o_coverage=**0**;

o_bcprev=**0**;
o_bcbias=**0**;
o_bcrelbias=**0**;
o_bcstbias=**0**;
o_bctvar=**0**;
o_bccoverage=**0**;

c_prev=**0**;
c_bias=**0**;
c_relbias=**0**;
c_stbias=**0**;
c_tvar=**0**;
c_coverage=**0**;

c_bcprev=**0**;
c_bcbias=**0**;
c_bcrelbias=**0**;
c_bcstbias=**0**;
c_bctvar=**0**;
c_bccoverage=**0**;

oc_prev=**0**;
oc_bias=**0**;
oc_relbias=**0**;
oc_stbias=**0**;
oc_tvar=**0**;

```
oc_coverage=0;

oc_bcprev=0;
oc_bcbias=0;
oc_bcrelbias=0;
oc_bcstbias=0;
oc_bctvar=0;
oc_bccoverage=0;

t_yimpute=j(ntot,nimpute,.);
t_bcyimpute=j(ntot,nimpute,.);
o_yimpute=j(ntot,nimpute,.); **O_YIMPUTE: imputed values of MI in frequentist with
observed disease status only**;
o_bcyimpute=j(ntot,nimpute,.);
c_yimpute=j(ntot,nimpute,.); **C_YIMPUTE: imputed values of MI in frequentist with imputed
covariates**;
c_bcyimpute=j(ntot,nimpute,.);
oc_yimpute=j(ntot,nimpute,.); **OC_YIMPUTE: imputed values of MI in frequentist with
observed disease status and covariates**;
oc_bcyimpute=j(ntot,nimpute,.);

**Main Body of Simulation**;
do r=1 to nsim;
**Generate a continuous covariate data and a binary covariate data**;
x1=j(ntot,1,.);
call randseed(seed1);
call randgen (x1,'normal',0,1);
x2=j(ntot,1,.);
call randseed(seed2);
call randgen (x2,'Bernoulli',.50);
**Create true disease status**;
do i=1 to ntot;
        pi=exp(beta0+beta1*x1[i]+beta2*x2[i])/(1+exp(beta0+beta1*x1[i]+beta2*x2[i]));
        call randseed(seed3);
        call randgen(yi,'bernoulli',pi);
        alldata[i,2]=yi;
end;
prev1=sum(alldata[,2])/ntot;
ytrue=alldata[,2];
prev=prev+prev1;
**Create observed disease status**;
do i=1 to ntot;
        underp=1/(1+exp(-(rho0+rho1*x1[i]+rho2*x2[i])));
        call randseed(seed4);
        call randgen(indobs,'bernoulli',underp); **Misclassification probability conditional on
true covariates**;
```

```
        if alldata[i,2]=1 then alldata[i,3]=indobs;
        else alldata[i,3]=0;
end;
obsprev1=sum(alldata[,3])/ntot;
obs_prev=obs_prev+obsprev1;
obs_bias1=prev1-obsprev1;
obs_bias=obs_bias+obs_bias1;

**Estimate the sensitivity and specificity of the diagnosis(U)**;
SN_count=0;
SP_count=0;
do i=1 to ntot;
        if alldata[i,2]=1 then do;
                if alldata[i,3]=1 then SN_count=SN_count+1;
        end;
        if alldata[i,2]=0 then do;
                if alldata[i,3]=0 then SP_count=SP_count+1;
        end;
end;
SN_U=SN_U+SN_count/sum(alldata[,2]);
SP_U=SP_U+SP_count/(ntot-sum(alldata[,2]));

**Create observed covariates**;
**Non-differential**;
m1=j(ntot,1,.); **The observed values of the continuous covariate**;
epsilon=j(ntot,1,.);
call randseed(seed5);
call randgen(epsilon,'normal',0,MEvar);
do i=1 to ntot;
        m1[i]=x1[i]+epsilon[i];
end;

m2=j(ntot,1,.); **The measurement of the dichotomous covariate which is assumed to be non-
differential**;
call randseed(seed6);
call randgen(indobs1,'bernoulli',SN);
call randseed(seed7);
call randgen(indobs2,'bernoulli',FP);
do i=1 to ntot;
        if x2[i]=1 then m2[i,1] = indobs1[i];
        else m2[i,1]= indobs2[i];
end;

**Create indicator for membership in the validation cohort and output validation cohort data into
a new dataset for analysis**;
**MCAR**;
```

```
call randseed(seed8);
call randgen(indval,'uniform');
validss=valrate*ntot;
temp=rank(indval);
**Create indicator for membership in the validation cohort and output validation cohort data into
a new dataset for analysis**;
**MAR**;
call randseed(seed8);
call randgen(indval,'uniform');
validss=valrate*ntot;
temp=rank(indval#(alldata[,3]+0.5));
**Create indicator for membership in the validation cohort and output validation cohort data into
a new dataset for analysis**;
**MNAR**;
call randseed(seed8);
call randgen(indval,'uniform');
validss=valrate*ntot;
temp=rank(indval#(alldata[,2]+0.5));
yset=j(validss,9,1);
count=0;
do i=1 to ntot;
        if temp[i]<=validss then do;
                alldata[i,1]=1;
                count=count+1;
                yset[count,1]=alldata[i,2];
                yset[count,2]=1;
                yset[count,3]=alldata[i,3];
                yset[count,4]=x1[i];
                yset[count,5]=x2[i];
                yset[count,7]=m1[i];
                yset[count,9]=m2[i];
        end;
        if temp[i]>validss then alldata[i,1]=0;
end;

**Impute the true covariates**;
call regress(yset[,4],yset[,{6 7}],phi,covphi);
call betagen(phimat,phi,covphi,nimpute,seed9); **Impute multiple coefficients for x1**;
ix1=j(ntot,nimpute,.);
do k=1 to nimpute;
        ix1_fill=j(ntot,1,.);
        do i=1 to ntot;
                if alldata[i,1]=1 then ix1_fill[i]=x1[i];
                else ix1_fill[i]=phimat[k,1]+phimat[k,2]*m1[i]; **Linear regression to predict
true covariates**;
        end;
```

```
        ix1[,k]=ix1_fill;
end;
call biascor(x2delta,x2covdelta,yset[,5],yset[,48],b);
call betagen(deltamat,x2delta,x2covdelta,nimpute,seed10);
ix2=j(ntot,nimpute,.);
do k=1 to nimpute;
        ix2_fill=j(ntot,1,.);
        do i=1 to ntot;
                if alldata[i,1]=0 then do;

        epx2=exp(deltamat[k,1]+deltamat[k,2]*m2[i])/(1+exp(deltamat[k,1]+deltamat[k,2]*m2[i
]));
                        call randseed(seed11);
                        call randgen(ex,'bernoulli',epx2);
                        ix2_fill[i]=ex;
                end;
                else if alldata[i,1]=1 then ix2_fill[i]=x2[i];
        end;
        ix2[,k]=ix2_fill;
end;

**Original with true covariates ie the disease model**;
call binest(t_beta,t_cov,yset[,1],yset[,{2 4 5}],b);
call betagen(tbetamat,t_beta,t_cov,nimpute,seed12);
t_storval=j(nimpute,1,.);
t_storvar=j(nimpute,1,.);
do k=1 to nimpute;
        t_prob=j(ntot,1,.);
        tfill_vec=j(ntot,1,.);
        t_indvar=j(ntot,1,.);
        do i=1 to ntot;
                if tbetamat[k,1]+tbetamat[k,2]*x1[i]+tbetamat[k,3]*x2[i]>0 then tp=1/(1+exp(-
(tbetamat[k,1]+tbetamat[k,2]*x1[i]+tbetamat[k,3]*x2[i])));
                else
tp=exp(tbetamat[k,1]+tbetamat[k,2]*x1[i]+tbetamat[k,3]*x2[i])/(1+exp(tbetamat[k,1]+tbetamat[
k,2]*x1[i]+tbetamat[k,3]*x2[i]));
                t_prob[i]=tp;
                if alldata[i,1]=0 then do;
                        call randseed(seed13,0);
                        call randgen(tfill,'bernoulli',t_prob[i]);
                        tfill_vec[i]=tfill;
                end;
                else if alldata[i,1]=1 then tfill_vec[i]=alldata[i,2];
        end;
        t_yimpute[,k]=tfill_vec;
        tprev=sum(t_yimpute[,k])/ntot;
```

```
        do i=1 to ntot;
                t_indvar[i]=t_prob[i]*(1-t_prob[i]);
        end;
        tpredvar=sum(t_indvar)/ntot**2;
        t_storval[k]=tprev;
        t_storvar[k]=tpredvar;
end;

t_prev0=sum(t_storval)/nimpute;
t_prev=t_prev+t_prev0;
t_bias0=t_prev0-prev1;
t_bias=t_bias+t_bias0;
t_relbias0=t_bias0/prev1; **The bias relative to the true prevelance**;
t_relbias=t_relbias+t_relbias0;
t_wvar=sum(t_storvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM based
on observed response only**;

do q=1 to nimpute;
        t_bvar=sum((t_storval[q]-sum(t_storval)/nimpute)**2)/(nimpute-1);
end;

t_tvar0=t_wvar+(1+1/nimpute)*t_bvar; **Total variability of Bias Correction MIM based on
observed response only**;
t_tvar=t_tvar+t_tvar0;
t_SE=sqrt(t_tvar0);
t_stbias0=t_bias0/t_SE;
t_stbias=t_stbias+t_stbias0;
t_LCI=t_prev0-1.96*t_SE;
t_UCI=t_prev0+1.96*t_SE;
ta=t_LCI<=prev1 & t_UCI>=prev1;
t_coverage=t_coverage+ta;

**Bias Correction with true covariates ie the disease model**;
call biascor(t_bcbeta,t_bccov,yset[,1],yset[,{2 4 5}],b);
call betagen(tbcbetamat,t_bcbeta,t_bccov,nimpute,seed14);
t_bcstorval=j(nimpute,1,.);
t_bcstorvar=j(nimpute,1,.);

do k=1 to nimpute;
        t_bcprob=j(ntot,1,.);
        tbcfill_vec=j(ntot,1,.);
        t_bcindvar=j(ntot,1,.);
        do i=1 to ntot;
                if tbcbetamat[k,1]+tbcbetamat[k,2]*x1[i]+tbcbetamat[k,3]*x2[i]>0 then
tbcp=1/(1+exp(-(tbcbetamat[k,1]+tbcbetamat[k,2]*x1[i]+tbcbetamat[k,3]*x2[i])));
```

```
                else
tbcp=exp(tbcbetamat[k,1]+tbcbetamat[k,2]*x1[i]+tbcbetamat[k,3]*x2[i])/(1+exp(tbcbetamat[k,1
]+tbcbetamat[k,2]*x1[i]+tbcbetamat[k,3]*x2[i]));
                t_bcprob[i]=tbcp;
                        if alldata[i,1]=0 then do;
                                call randseed(seed15,0);
                                call randgen(tbcfill,'bernoulli',t_bcprob[i]);
                                tbcfill_vec[i]=tbcfill;
                        end;
                        else if alldata[i,1]=1 then tbcfill_vec[i]=alldata[i,2];
        end;
        t_bcyimpute[,k]=tbcfill_vec;
        tbcprev=sum(t_bcyimpute[,k])/ntot;
        do i=1 to ntot;
                t_bcindvar[i]=t_bcprob[i]*(1-t_bcprob[i]);
        end;
        tbcpredvar=sum(t_bcindvar)/ntot**2;
        t_bcstorval[k]=tbcprev;
        t_bcstorvar[k]=tbcpredvar;
end;

t_bcprev0=sum(t_bcstorval)/nimpute;
t_bcprev=t_bcprev+t_bcprev0;
t_bcbias0=t_bcprev0-prev1;
t_bcbias=t_bcbias+t_bcbias0;
t_bcrelbias0=t_bcbias0/prev1; **The bias relative to the true prevelance**;
t_bcrelbias=t_bcrelbias+t_bcrelbias0;
t_bcwvar=sum(t_bcstorvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM
based on observed response only**;

do q=1 to nimpute;
        t_bcbvar=sum((t_bcstorval[q]-sum(t_bcstorval)/nimpute)**2)/(nimpute-1);
end;

t_bctvar0=t_bcwvar+(1+1/nimpute)*t_bcbvar; **Total variability of Bias Correction MIM based
on observed response only**;
t_bctvar=t_bctvar+t_bctvar0;

t_bcSE=sqrt(t_bctvar0);
t_bcstbias0=t_bcbias0/t_bcSE;
t_bcstbias=t_bcstbias+t_bcstbias0;
t_bcLCI=t_bcprev0-1.96*t_bcSE;
t_bcUCI=t_bcprev0+1.96*t_bcSE;
tbca=t_bcLCI<=prev1 & t_bcUCI>=prev1;
t_bccoverage=t_bccoverage+tbca;
```

```
**Original only with observed response**;
call binest(obeta,ocov,yset[,1],yset[,2:3],b);
call betagen(obetamat,obeta,ocov,nimpute,seed16);
o_storval=j(nimpute,1,.);
o_storvar=j(nimpute,1,.);

do k=1 to nimpute;
	o_prob=j(ntot,1,.);
	ofill_vec=j(ntot,1,.);
	o_indvar=j(ntot,1,.);
	do i=1 to ntot;
		if obetamat[k,1]+obetamat[k,2]*alldata[i,3]>0 then op=1/(1+exp(-
(obetamat[k,1]+obetamat[k,2]*alldata[i,3])));
		else
op=exp(obetamat[k,1]+obetamat[k,2]*alldata[i,3])/(1+exp(obetamat[k,1]+obetamat[k,2]*alldata[
i,3]));
		o_prob[i]=op;
			if alldata[i,1]=0 then do;
				call randseed(seed17);
				call randgen(o_fill,'bernoulli',o_prob[i]);
				ofill_vec[i]=o_fill;
			end;
			else if alldata[i,1]=1 then ofill_vec[i]=alldata[i,2];
	end;
	o_yimpute[,k]=ofill_vec;
	oprev=sum(o_yimpute[,k])/ntot;
	do i=1 to ntot;
		o_indvar[i]=o_prob[i]*(1-o_prob[i]);
	end;
	opredvar=sum(o_indvar)/ntot**2;
	o_storval[k]=oprev;
	o_storvar[k]=opredvar;
end;

o_prev0=sum(o_storval)/nimpute;
o_prev=o_prev+o_prev0;
o_bias0=o_prev0-prev1;
o_bias=o_bias+o_bias0;

o_relbias0=o_bias0/prev1; **The bias relative to the true prevelance**;
o_relbias=o_relbias+o_relbias0;
o_wvar=sum(o_storvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM based
on observed response only**;

do q=1 to nimpute;
	o_bvar=sum((o_storval[q]-sum(o_storval)/nimpute)**2)/(nimpute-1);
```

```
end;

o_tvar0=o_wvar+(1+1/nimpute)*o_bvar; **Total variability of Bias Correction MIM based on
observed response only**;
o_tvar=o_tvar+o_tvar0;
o_SE=sqrt(o_tvar0);
o_stbias0=o_bias0/o_SE;
o_stbias=o_stbias+o_stbias0;
o_LCI=o_prev0-1.96*o_SE;
o_UCI=o_prev0+1.96*o_SE;
oa=o_LCI<=prev1 & prev1<=o_UCI;
o_coverage=o_coverage+oa;

**Bias Correction only with observed response**;
call biascor(obcbeta,obccov,yset[,1],yset[,2:3],b);
call betagen(obcbetamat,obcbeta,obccov,nimpute,seed18);
o_bcstorval=j(nimpute,1,.);
o_bcstorvar=j(nimpute,1,.);

do k=1 to nimpute;
        o_bcprob=j(ntot,1,.);
        obcfill_vec=j(ntot,1,.);
        o_bcindvar=j(ntot,1,.);
        do i=1 to ntot;
                if obcbetamat[k,1]+obcbetamat[k,2]*alldata[i,3]>0 then obcp=1/(1+exp(-
(obcbetamat[k,1]+obcbetamat[k,2]*alldata[i,3])));
                else
obcp=exp(obcbetamat[k,1]+obcbetamat[k,2]*alldata[i,3])/(1+exp(obcbetamat[k,1]+obcbetamat[
k,2]*alldata[i,3]));
                o_bcprob[i]=obcp;
                if alldata[i,1]=0 then do;
                        call randseed(seed19);
                        call randgen(o_bcfill,'bernoulli',o_bcprob[i]);
                        obcfill_vec[i]=o_bcfill;
                end;
                else if alldata[i,1]=1 then obcfill_vec[i]=alldata[i,2];
        end;
        o_bcyimpute[,k]=obcfill_vec;
        obcprev=sum(o_bcyimpute[,k])/ntot;
        do i=1 to ntot;
                o_bcindvar[i]=o_bcprob[i]*(1-o_bcprob[i]);
        end;
        obcpredvar=sum(o_bcindvar)/ntot**2;
        o_bcstorval[k]=obcprev;
        o_bcstorvar[k]=obcpredvar;
end;
```

```
o_bcprev0=sum(o_bcstorval)/nimpute;
o_bcprev=o_bcprev+o_bcprev0;
o_bcbias0=o_bcprev0-prev1;
o_bcbias=o_bcbias+o_bcbias0;
o_bcrelbias0=o_bcbias0/prev1; **The bias relative to the true prevelance**;
o_bcrelbias=o_bcrelbias+o_bcrelbias0;
o_bcwvar=sum(o_bcstorvar)/nimpute;  **Within Imputation Variance of Bias Correction MIM
based on observed response only**;

do q=1 to nimpute;
        o_bcbvar=sum((o_bcstorval[q]-sum(o_bcstorval)/nimpute)**2)/(nimpute-1);
end;

o_bctvar0=o_bcwvar+(1+1/nimpute)*o_bcbvar; **Total variability of Bias Correction MIM
based on observed response only**;
o_bctvar=o_bctvar+o_bctvar0;
o_bcSE=sqrt(o_bctvar0);
o_bcstbias0=o_bcbias0/o_bcSE;
o_bcstbias=o_bcstbias+o_bcstbias0;
o_bcLCI=o_bcprev0-1.96*o_bcSE;
o_bcUCI=o_bcprev0+1.96*o_bcSE;
obca=o_bcLCI<=prev1 & prev1<=o_bcUCI;
o_bccoverage=o_bccoverage+obca;

**Original with imputed covariates**;
call binest(cbeta,ccov,yset[,1],yset[,{2 4 5}],b);
call betagen(cbetamat,cbeta,ccov,nimpute,seed20);
c_storval=j(nimpute,1,.);
c_storvar=j(nimpute,1,.);

do k=1 to nimpute;
        c_prob=j(ntot,1,.);
        cfill_vec=j(ntot,1,.);
        c_indvar=j(ntot,1,.);
        do i=1 to ntot;
                if cbetamat[k,1]+cbetamat[k,2]*ix1[i,k]+cbetamat[k,3]*ix2[i,k]>0 then
cp=1/(1+exp(-(cbetamat[k,1]+cbetamat[k,2]*ix1[i,k]+cbetamat[k,3]*ix2[i,k])));
                else
cp=exp(cbetamat[k,1]+cbetamat[k,2]*ix1[i,k]+cbetamat[k,3]*ix2[i,k])/(1+exp(cbetamat[k,1]+cb
etamat[k,2]*ix1[i,k]+cbetamat[k,3]*ix2[i,k]));
                c_prob[i]=cp;
                if alldata[i,1]=0 then do;
                        call randseed(seed21,0);
                        call randgen(cfill,'bernoulli',c_prob[i]);
                        cfill_vec[i]=cfill;
```

```
                end;
                else if alldata[i,1]=1 then cfill_vec[i]=alldata[i,2];
        end;
        c_yimpute[,k]=cfill_vec;
        cprev=sum(c_yimpute[,k])/ntot;
        do i=1 to ntot;
                c_indvar[i]=c_prob[i]*(1-c_prob[i]);
        end;
        cpredvar=sum(c_indvar)/ntot**2;
        c_storval[k]=cprev;
        c_storvar[k]=cpredvar;
end;

c_prev0=sum(c_storval)/nimpute;
c_prev=c_prev+c_prev0;
c_bias0=c_prev0-prev1;
c_bias=c_bias+c_bias0;
c_relbias0=c_bias0/prev1;
c_relbias=c_relbias+c_relbias0;
c_wvar=sum(c_storvar)/nimpute; **Within Imputation Variance of Bias Correction based on
observed response and covariates**;

do q=1 to nimpute;
        c_bvar=sum((c_storval[q]-sum(c_storval)/nimpute)**2)/(nimpute-1);
end;

c_tvar0=c_wvar+(1+1/nimpute)*c_bvar; **Total Variability of Bias Correction based on
observed response and covariates**;
c_tvar=c_tvar+c_tvar0;
c_SE=sqrt(c_tvar0);
c_stbias0=c_bias0/c_SE;
c_stbias=c_stbias+c_stbias0;
c_LCI=c_prev0-1.96*c_SE;
c_UCI=c_prev0+1.96*c_SE;
ca=c_LCI<=prev1 & prev1<=c_UCI;
c_coverage=c_coverage+ca;

**Bias Correction with imputed covariates**;
call biascor(cbc_beta,cbc_cov,yset[,1],yset[,{2 4 5}],b);
call betagen(cbcbetamat,cbc_beta,cbc_cov,nimpute,seed22);
c_bcstorval=j(nimpute,1,.);
c_bcstorvar=j(nimpute,1,.);

do k=1 to nimpute;
        c_bcprob=j(ntot,1,.);
        cbcfill_vec=j(ntot,1,.);
```

```
        c_bcindvar=j(ntot,1,.);
        do i=1 to ntot;
                if cbcbetamat[k,1]+cbcbetamat[k,2]*ix1[i,k]+cbcbetamat[k,3]*ix2[i,k]>0 then
cbcp=1/(1+exp(-(cbcbetamat[k,1]+cbcbetamat[k,2]*ix1[i,k]+cbcbetamat[k,3]*ix2[i,k])));
                else
cbcp=exp(cbcbetamat[k,1]+cbcbetamat[k,2]*ix1[i,k]+cbcbetamat[k,3]*ix2[i,k])/(1+exp(cbcbeta
mat[k,1]+cbcbetamat[k,2]*ix1[i,k]+cbcbetamat[k,3]*ix2[i,k]));
                c_bcprob[i]=cbcp;
                if alldata[i,1]=0 then do;
                        call randseed(seed23,0);
                        call randgen(cbcfill,'bernoulli',c_bcprob[i]);
                        cbcfill_vec[i]=cbcfill;
                end;
                else if alldata[i,1]=1 then cfill_vec[i]=alldata[i,2];
        end;
        c_bcyimpute[,k]=cbcfill_vec;
        cbcprev=sum(c_bcyimpute[,k])/ntot;
        do i=1 to ntot;
                c_bcindvar[i]=c_bcprob[i]*(1-c_bcprob[i]);
        end;
        cbcpredvar=sum(c_bcindvar)/ntot**2;
        c_bcstorval[k]=cbcprev;
        c_bcstorvar[k]=cbcpredvar;
end;

c_bcprev0=sum(c_bcstorval)/nimpute;
c_bcprev=c_bcprev+c_bcprev0;
c_bcbias0=c_bcprev0-prev1;
c_bcbias=c_bcbias+c_bcbias0;
c_bcrelbias0=c_bcbias0/prev1;
c_bcrelbias=c_bcrelbias+c_bcrelbias0;
c_bcwvar=sum(c_bcstorvar)/nimpute; **Within Imputation Variance of Bias Correction based
on observed response and covariates**;

do q=1 to nimpute;
        c_bcbvar=sum((c_bcstorval[q]-sum(c_bcstorval)/nimpute)**2)/(nimpute-1);
end;

c_bctvar0=c_bcwvar+(1+1/nimpute)*c_bcbvar; **Total Variability of Bias Correction based on
observed response and covariates**;
c_bctvar=c_bctvar+c_bctvar0;
c_bcSE=sqrt(c_bctvar0);
c_bcstbias0=c_bcbias0/c_bcSE;
c_bcstbias=c_bcstbias+c_bcstbias0;
c_bcLCI=c_bcprev0-1.96*c_bcSE;
c_bcUCI=c_bcprev0+1.96*c_bcSE;
```

```
cbca=c_bcLCI<=prev1 & prev1<=c_bcUCI;
c_bccoverage=c_bccoverage+cbca;

**Original with observed response and imputed covariates**;
call binest(ocbeta,occov,yset[,1],yset[,2:5],b);
call betagen(ocbetamat,ocbeta,occov,nimpute,seed24);
oc_storval=j(nimpute,1,.);
oc_storvar=j(nimpute,1,.);

do k=1 to nimpute;
	oc_prob=j(ntot,1,.);
	ocfill_vec=j(ntot,1,.);
	oc_indvar=j(ntot,1,.);
	do i=1 to ntot;
		if
ocbetamat[k,1]+ocbetamat[k,2]*alldata[i,3]+ocbetamat[k,3]*ix1[i,k]+ocbetamat[k,4]*ix2[i,k]>0
then ocpv=1/(1+exp(-
(ocbetamat[k,1]+ocbetamat[k,2]*alldata[i,3]+ocbetamat[k,3]*ix1[i,k]+ocbetamat[k,4]*ix2[i,k])));
		else
ocpv=exp(ocbetamat[k,1]+ocbetamat[k,2]*alldata[i,3]+ocbetamat[k,3]*ix1[i,k]+ocbetamat[k,4]*
ix2[i,k])/(1+exp(ocbetamat[k,1]+ocbetamat[k,2]*alldata[i,3]+ocbetamat[k,3]*ix1[i,k]+ocbetamat
[k,4]*ix2[i,k]));
		oc_prob[i]=ocpv;
		if alldata[i,1]=0 then do;
			call randseed(seed25,0);
			call randgen(ocfill,'bernoulli',oc_prob[i]);
			ocfill_vec[i]=ocfill;
		end;
		else if alldata[i,1]=1 then ocfill_vec[i]=alldata[i,2];
	end;
	oc_yimpute[,k]=ocfill_vec;
	ocpredprev=sum(oc_yimpute[,k])/ntot;
	do i=1 to ntot;
		oc_indvar[i]=oc_prob[i]*(1-oc_prob[i]);
	end;
	ocpredvar=sum(oc_indvar)/ntot**2;
	oc_storval[k]=ocpredprev;
	oc_storvar[k]=ocpredvar;
end;

oc_prev0=(sum(oc_storval)/nimpute);
oc_prev=oc_prev+oc_prev0;
oc_bias0=oc_prev0-prev1;
oc_bias=oc_bias+oc_bias0;
oc_relbias0=oc_bias0/prev1;
oc_relbias=oc_relbias+oc_relbias0;
```

```
oc_wvar=sum(oc_storvar)/nimpute; **Within Imputation of logistic MIM based on observed
response and covariates**;

do q=1 to nimpute;
        oc_bvar=sum((oc_storval[q]-sum(oc_storval)/nimpute)**2)/(nimpute-1);
end;

oc_tvar0=oc_wvar+(1+1/nimpute)*oc_bvar; **Total variability of logistic MIM based on
observed response and imputed covariates**;
oc_tvar=oc_tvar+oc_tvar0;
oc_SE=sqrt(oc_tvar0);
oc_stbias0=oc_bias0/oc_SE;
oc_stbias=oc_stbias+oc_stbias0;
oc_LCI=oc_prev0-1.96*oc_SE;
oc_UCI=oc_prev0+1.96*oc_SE;
oca=oc_LCI<=prev1 & prev1<=oc_UCI;
oc_coverage=oc_coverage+oca;

**Bias Correction with observed response and imputed covariates**;
call biascor(ocbcbeta,ocbccov,yset[,1],yset[,2:5],b);
call betagen(ocbcbetamat,ocbcbeta,ocbccov,nimpute,seed26);
oc_bcstorval=j(nimpute,1,.);
oc_bcstorvar=j(nimpute,1,.);

do k=1 to nimpute;
        oc_bcprob=j(ntot,1,.);
        ocbcfill_vec=j(ntot,1,.);
        oc_bcindvar=j(ntot,1,.);
        do i=1 to ntot;
                if
ocbcbetamat[k,1]+ocbcbetamat[k,2]*alldata[i,3]+ocbcbetamat[k,3]*ix1[i,k]+ocbcbetamat[k,4]*i
x2[i,k]>0 then ocbcpv=1/(1+exp(-
(ocbcbetamat[k,1]+ocbcbetamat[k,2]*alldata[i,3]+ocbcbetamat[k,3]*ix1[i,k]+ocbcbetamat[k,4]*
ix2[i,k])));
                else
ocbcpv=exp(ocbcbetamat[k,1]+ocbcbetamat[k,2]*alldata[i,3]+ocbcbetamat[k,3]*ix1[i,k]+ocbcbe
tamat[k,4]*ix2[i,k])/(1+exp(ocbcbetamat[k,1]+ocbcbetamat[k,2]*alldata[i,3]+ocbcbetamat[k,3]*
ix1[i,k]+ocbcbetamat[k,4]*ix2[i,k]));
                oc_bcprob[i]=ocbcpv;
                if alldata[i,1]=0 then do;
                        call randseed(seed27,0);
                        call randgen(ocbcfill,'bernoulli',oc_bcprob[i]);
                        ocbcfill_vec[i]=ocbcfill;
                end;
                else if alldata[i,1]=1 then ocbcfill_vec[i]=alldata[i,2];
        end;
```

```
        oc_bcyimpute[,k]=ocbcfill_vec;
        ocbcpredprev=sum(oc_bcyimpute[,k])/ntot;
        do i=1 to ntot;
                oc_bcindvar[i]=oc_bcprob[i]*(1-oc_bcprob[i]);
        end;
        ocbcpredvar=sum(oc_bcindvar)/ntot**2;
        oc_bcstorval[k]=ocbcpredprev;
        oc_bcstorvar[k]=ocbcpredvar;
end;

oc_bcprev0=(sum(oc_bcstorval)/nimpute);
oc_bcprev=oc_bcprev+oc_bcprev0;
oc_bcbias0=oc_bcprev0-prev1;
oc_bcbias=oc_bcbias+oc_bcbias0;
oc_bcrelbias0=oc_bcbias0/prev1;
oc_bcrelbias=oc_bcrelbias+oc_bcrelbias0;
oc_bcwvar=sum(oc_bcstorvar)/nimpute; **Within Imputation of logistic MIM based on
observed response and covariates**;

do q=1 to nimpute;
        oc_bcbvar=sum((oc_bcstorval[q]-sum(oc_bcstorval)/nimpute)**2)/(nimpute-1);
end;

oc_bctvar0=oc_bcwvar+(1+1/nimpute)*oc_bcbvar; **Total variability of logistic MIM based on
observed response and imputed covariates**;
oc_bctvar=oc_bctvar+oc_bctvar0;
oc_bcSE=sqrt(oc_bctvar0);
oc_bcstbias0=oc_bcbias0/oc_bcSE;
oc_bcstbias=oc_bcstbias+oc_bcstbias0;
oc_bcLCI=oc_bcprev0-1.96*oc_bcSE;
oc_bcUCI=oc_bcprev0+1.96*oc_bcSE;
ocbca=oc_bcLCI<=prev1 & prev1<=oc_bcUCI;
oc_bccoverage=oc_bccoverage+ocbca;
end; **END of the simulation**;

prev=prev/nsim;
obs_prev=obs_prev/nsim;
obs_bias=obs_bias/nsim;
SN_U=SN_U/nsim;
SP_U=SP_U/nsim;

t_prev=t_prev/nsim;
t_bias=t_bias/nsim;
t_relbias=t_relbias/nsim;
t_stbias=t_stbias/nsim;
t_tvar=t_tvar/nsim;
```

```
t_mse=t_tvar+t_bias**2;
t_rmse=sqrt(t_mse);
t_coverage=t_coverage/nsim;

t_bcprev=t_bcprev/nsim;
t_bcbias=t_bcbias/nsim;
t_bcrelbias=t_bcrelbias/nsim;
t_bcstbias=t_bcstbias/nsim;
t_bctvar=t_bctvar/nsim;
t_bcmse=t_bctvar+t_bcbias**2;
t_bcrmse=sqrt(t_bcmse);
t_bccoverage=t_bccoverage/nsim;

o_prev=o_prev/nsim;
o_bias=o_bias/nsim;
o_relbias=o_relbias/nsim;
o_stbias=o_stbias/nsim;
o_tvar=o_tvar/nsim;
o_mse=o_tvar+o_bias**2;
o_rmse=sqrt(o_mse);
o_coverage=o_coverage/nsim;

o_bcprev=o_bcprev/nsim;
o_bcbias=o_bcbias/nsim;
o_bcrelbias=o_bcrelbias/nsim;
o_bcstbias=o_bcstbias/nsim;
o_bctvar=o_bctvar/nsim;
o_bcmse=o_bctvar+o_bcbias**2;
o_bcrmse=sqrt(o_bcmse);
o_bccoverage=o_bccoverage/nsim;

c_prev=c_prev/nsim;
c_bias=c_bias/nsim;
c_relbias=c_relbias/nsim;
c_stbias=c_stbias/nsim;
c_tvar=c_tvar/nsim;
c_mse=c_tvar+c_bias**2;
c_rmse=sqrt(c_mse);
c_coverage=c_coverage/nsim;

c_bcprev=c_bcprev/nsim;
c_bcbias=c_bcbias/nsim;
c_bcrelbias=c_bcrelbias/nsim;
c_bcstbias=c_bcstbias/nsim;
c_bctvar=c_bctvar/nsim;
c_bcmse=c_bctvar+c_bcbias**2;
```

```
c_bcrmse=sqrt(c_bcmse);
c_bccoverage=c_bccoverage/nsim;

oc_prev=oc_prev/nsim;
oc_bias=oc_bias/nsim;
oc_relbias=oc_relbias/nsim;
oc_stbias=oc_stbias/nsim;
oc_tvar=oc_tvar/nsim;
oc_mse=oc_tvar+oc_bias**2;
oc_rmse=sqrt(oc_mse);
oc_coverage=oc_coverage/nsim;

oc_bcprev=oc_bcprev/nsim;
oc_bcbias=oc_bcbias/nsim;
oc_bcrelbias=oc_bcrelbias/nsim;
oc_bcstbias=oc_bcstbias/nsim;
oc_bctvar=oc_bctvar/nsim;
oc_bcmse=oc_bctvar+oc_bcbias**2;
oc_bcrmse=sqrt(oc_bcmse);
oc_bccoverage=oc_bccoverage/nsim;

Parameter=ntot||sensitivity||valrate||nimpute||MEvar||SN||FP;
OBS=nsim||prev||obs_prev||obs_bias||SN_U||SP_U;
TMIM=t_prev||t_bias||t_relbias||t_stbias||t_tvar||t_rmse||t_coverage;
TBCMIM=t_bcprev||t_bcbias||t_bcrelbias||t_bcstbias||t_bctvar||t_bcrmse||t_bccoverage;
OMIM=o_prev||o_bias||o_relbias||o_stbias||o_tvar||o_rmse||o_coverage;
OBCMIM=o_bcprev||o_bcbias||o_bcrelbias||o_bcstbias||o_bctvar||o_bcrmse||o_bccoverage;
CMIM=c_prev||c_bias||c_relbias||c_stbias||c_tvar||c_rmse||c_coverage;
CBCMIM=c_bcprev||c_bcbias||c_bcrelbias||c_bcstbias||c_bctvar||c_bcrmse||c_bccoverage;
OCMIM=oc_prev||oc_bias||oc_relbias||oc_stbias||oc_tvar||oc_rmse||oc_coverage;
OCBCMIM=oc_bcprev||oc_bcbias||oc_bcrelbias||oc_bcstbias||oc_bctvar||oc_bcrmse||oc_bccover
age;
Result1=Parameter||OBS||TMIM||TBCMIM||OMIM||OBCMIM||CMIM||CBCMIM||OCMIM||OC
BCMIM;

ResultAll[1#(MEvar=1)+13#(MEvar=2)+0#(SN=0.9)+6#(SN=0.7)+0#(nimpute=3)+1#(nimpute
=5)+2#(nimpute=10)+3#(nimpute=15)+4#(nimpute=20)+5#(nimpute=40),]=Result1;

Varnames={TotalNumbers Sensitivity ValRate NImpute AddVar XSensitivity XFalsePositive
SimulationNumber Prevalence ObservedPrevalence ObservedPrevBias TestUSensitiviy
TestUSpecificity PrevT BiasT RelBiasT StBiasT TotVarT RMSET CoverageT PrevTBC
BiasTBC RelBiasTBC StBiasTBC TotVarTBC RMSETBC CoverageTBC PrevO BiasO
RelBiasO StBiasO TotVarO RMSEO CoverageO PrevOBC BiasOBC RelBiasOBC StBiasOBC
TotVarOBC RMSEOBC CoverageOBC PrevC BiasC RelBiasC StBiasC TotVarC RMSEC
CoverageC PrevCBC BiasCBC RelBiasCBC StBiasCBC TotVarCBC RMSECBC
CoverageCBC PrevOC BiasOC RelBiasOC StBiasOC TotVarOC RMSEOC CoverageOC
```

```
PrevOCBC BiasOCBC RelBiasOCBC StBiasOCBC TotVarOCBC RMSEOCBC
CoverageOCBC};

end; **END of number of imputation loop**;
end; **END of misclassification characteristics**;
end; **END of measurement error characteristics**;

create MCAR2.MCAR2P05S65 from ResultAll [colname=varnames];
append from ResultAll;
close MCAR2.MCAR2P05S65;

SEED=seed1||seed2||seed3||seed4||seed5||seed6||seed7||seed8||seed9||seed10||seed11||seed12||seed1
3||seed14||seed15||seed16||seed17||seed18||seed19||seed20||seed21||seed22||seed23||seed24||seed25|
|seed26||seed27;

create MCAR2.SEED2P05S65 from SEED;
append from SEED;
close MCAR2.SEED2P05S65;

quit;
```